Grapholinguistics
in the 21st Century—2020

/gʁafematik/
Proceedings

June 17-19, 2020
Yannis Haralambous (Ed.)

Part II

Grapholinguistics and Its Applications   5

Series Editor

Yannis Haralambous (Ed.)

Grapholinguistics in the 21st Century

/gʁafematik/
June 15–17, 2020 (online)
Proceedings   Part II

Fluxus Editions

Cover illustration: *Weapon*, in heptapod B language, courtesy of Wolfram Companies (https://github.com/WolframResearch/Arrival-Movie-Live-Coding). From the movie *Arrival* (2016) by Denis Villeneuve, inspired by the short story Ted Chiang, "Story of Your Life," *Starflight*, Vol. 2, New York: Tor Books, 1998.

Table of Contents

PART II

Preface

The second edition of the *Grapholinguistics in the 21st Century* conference was held online, owing to the COVID-19 pandemic, on June 15–17, 2020. In these *Proceedings* are collected forty-two contributions derived from oral or poster presentations.

The first five papers (Neef, Evertz-Rittich, Osterkamp & Schreiber, Presutti, and Gnanadesikan) contribute to the theoretical body of grapholinguistics, addressing core concepts: the written utterance, the written word, phonography and morphography, the interdependence of speech and writing, and the native script effect. Offering a global perspective, the paper by Meletis, author of the recently released *The nature of writing: A theory of grapholinguistics*, discusses the activity of being a grapholinguist, its challenges and promises.

The common theme of the papers by Salomon, Harbour, and Elti di Rodeano is *beginnings*: script creation or transfer (inspired by the runic script); the influence of grammar on writing system evolution and the birth of vowels; transmission of the first alphabets.

The next block of six papers deals with (typo)graphetics: Véry explores textual space; Haralambous, Landragin & Handa study graphemic and graphetic methods in speculative fiction; Wachendorff examines urban spaces in the Ruhr area; Kulish gives a survey of nonstandard, "emotional," punctuation; Bergergausen & Huot-Marchand and Pierson present their font creation projects, respectively, "Missing scripts" and "PIM" (ancient monetary inscriptions).

In the papers that follow, Neuman gives an account of spelling variation in Modern Hebrew from a sociocultural point of view; Dürscheid provides us with insight on the use of emojis in social media; Melka & Schoch investigate the possibility of communication, be it visual or auditory, with unknown intelligence/s.

The last paper of the first part of the *Proceedings* provides an artist's perspective: Kettaneh gives us an account on her very inspired work involving written language in many forms.

The second part of the *Proceedings* starts with a block of four papers in the area of sinographemics: Joyce & Masuda explore three-character and four-character words in Japanese; Honda provides us with a modular-theoretic approach to the Japanese writing system; Myers and Morioka deal with the internal structure of sinographs.

A group of eight contributions of historical nature follows. Stojanov deals with the description of punctuation in Western grammar books; Drozhashchikh, Efimova & Meshcheryakova with form-meaning regularities in Old English; Presutti with graphemics of new Romance phonemes in Italian; Fendel with Coptic alphabets; Giunashvili with Old Aramaic script in Georgia; Fedorova with Aztec emblems; Kelly with the Bougainville Naasioi Otomaung alphabet; Schoch & Melka with the Easter Island rongorongo script.

The next block of five papers deals with applications of the computer in grapholinguistics: Jee, Tamariz & Shillcock study sound-graphic systematicity in various fonts; Sheng, Colin & Perono Cacciafoco attempt to decipher Linear A by a brute force attack; Salgarella & Castellan present a palæographical database for Linear A; Donnelly describes a system for digitizing Swahili in Arabic script; Xu presents a semantic index for the Dongba script of the Naxi people of Southwest China.

Speech and writing are not the only modalities of languages. There is also gestuality, used in sign languages. Two papers deal with the written transcription of sign languages: Danet *et al.* present the TYPANNOT system; Bianchini discusses metalinguistic awareness. Among the authors of Danet *et al.* is also Dominique Boutet who succumbed to the COVID-19 disease a few weeks before the conference.

The three papers that follow deal with the confrontation of two scripts. Koch investigates that between Roman and Cyrillic for the Moldovan language; Awad, Mourad & Elamil study the use of punctuation in French-to-Arabic translation; Rashwan investigates the visual untranslatability of the Ancient Egyptian and Arabic writing systems.

The volume concludes with a supernatural touch, as Küster leads us in a tour of magical writing, from cuneiform acrostics to modern manga.

The volumetry of these *Proceedings* is important: its 42 papers were written by 62 authors, span 1,122 pages (an average of 26.8 pages per paper, with a maximum of 102 and a minimum of 12 pages) and contain 412 figures and 1,940 bibliographical references; the index stretches to 1,247 entries. For technical reasons, the printed version of the *Proceedings* has been split into two parts: Part I, from Neef to Kettaneh (pages 1 to 577) and Part II, from Joyce & Masuda to Küster (pages 579 to 1122). Both front matter (preface, table of contents, list of participants) and back matter (index) are provided in both parts, the former in Roman page numbering (i–xii) and the latter in Greek page numbering (α′–κγ′). Some papers use different illustrations and text styles for the printed black & white version and the online color version.

All presentations at the *Grapholinguistics in the 21st Century* 2020 conference were recorded and can be viewed on Youtube. The links can be found on the conference webpage (`https://grafematik2020.sciencesconf.org/` or `https://perma.cc/3TJ6-RCJ5`).

List of Participants
at the
Grapholinguistics in the 21st Century 2020
Conference

Ahlberg, Aija Katriina
Almuzaini, Rawan
Anderson, Debbie
Ashourinia, Kaveh
Avni, Shani
Awad, Dana
Ayre, Christine
Baggio, Pietro
Beaujard, Laurence
Bergerhausen, Johannes
Berning, Bianca
Bianchini, Claudia S.
Birk, Elisabeth
Březina, David
Caren, Daniel
Castellan, Simon
Chalkley, Kendra
Chew, Patrick
Clifton, John
Contesse, Adrien
Cooley, Christopher
Crellin, Robert
Dalma, Véry
Danet, Claire
Danziger, Eve
Davison, Phil
Donnelly, Kevin
Drozhashchikh, Nataliia
Dunlavey, Nicholas
Dürscheid, Christa
Dürst, Martin
Efimova, Elena
Elti di Rodeano, Sveva
Elvira Astoreca, Natalia

Esfahbod, Behnam
Evertz-Rittich, Martin
Farrando, Pere
Fedorova, Liudmila
Fendel, Victoria Beatrix
Fisher, Filipa
Folaron, Debbie
Gardner, William
Gaultney, Victor
Gautier, Antoine
Giunashvili, Helen
Gnanadesikan, Amalia
Gomez-Jimenez, Eva
Handa, Kenichi
Handel, Zev
Haralambous, Yannis
Harbour, Daniel
Honda, Keisuke
Huot-Marchand, Thomas
Hutto, Megan
Iannucci, David
Ikeda, Elissa
Issele, Joanna
Iyengar, Arvind
Izadpanah, Borna
Jee, Hana
Joyce, Terry
Judson, Anna
Kang, Michelle
Karakılçık, Pınar
Kelly, Piers
Kettaneh, Christine
Kobayashi-Better, Daniel
Koch, Christian

Kučera, Jan
Kulish, Olga
Küster, Marc
Landragin, Frédéric
Mansour, Kamal
Martinod, Emmanuella
Meilleur, Maurice
Meletis, Dimitrios
Melka, Tomi S.
Mescheryakova, Evgenia
Morioka, Tomohiko
Myers, James
Mykolaiv, Ivan
Nicolaou, Andrew
Osborn, J.R.
Osterkamp, Sven
Papazian, Hrant
Penney, Laurence
Perono Cacciafoco, Francesco
Pierson, Morgane
Plesniarski, Allan
Poth, Christina
Presutti, Stefano
Rashdan, Amany
Rashwan, Hany
Rauff, James
Riggs, Tamye
Roux, Élie
Sadan, Meir

Salgarella, Ester
Salomon, Corinna
Schindelin, Cornelia
Schneider, Elena
Schoch, Robert M.
Schönecker, Anna-Lisa
Schreiber, Gordian
Selvelli, Giustina
Serra, Laura
Shingledecker, Anna
Steele, Pippa
Stojanov, Tomislav
Taha, Haitham
Tauber, James
Tenenbaum, Abi
Ugray, Gábor
Verheijen, Lieke
Vlachou, Irene
Vranas, Apostolos
Wachendorff, Irmi
Waldispühl, Michelle
Walther, Urs
Wang, Ruoan
Wiebe, Bruce
Williamson, Ryan
Wöhrmann, Frithjof
Xu, Duoduo
Yang, Ben
Ziegler, Monica

Constructing Databases
of Japanese Three- and Four-Kanji
Compound Words

Some Observations Concerning
Their Morphological Structures

Terry Joyce · Hisashi Masuda

Abstract. As the principal component of the multiple script Japanese writing system (JWS), morphographic kanji function as the core units of graphematic representation for a considerable proportion of the Japanese lexicon (Joyce and Masuda, 2018; 2019; Joyce, Masuda, and Ogawa, 2014; Kobayashi, Yamashita, and Kageyama, 2016; Nomura, 1975; 1988). Deeply entwined with the morphographic nature of Japanese kanji (Joyce, 2011), the Japanese language offers especially fascinating opportunities for both linguistic and psycholinguistic investigations of compound words (Joyce, 2002; 2004; Masuda and Joyce, 2018). As contributions to the ongoing construction of a larger database project of Japanese lexical properties (Joyce, Hodošček, and Masuda, 2017; Joyce, Masuda, and Ogawa, 2014), which aims to facilitate such investigations in terms of experimental designs and stimuli preparation, this paper reports on two new database components for three-kanji (3KCWs) and four-kanji compound words (4KCWs) respectively. More specifically, the paper focuses on the results of analyzing their morphological structures. In contrast to 3KCWs, where the dominant morphological structure is attaching suffixes to existing two-kanji compound words (2KCWs), such as 可能性 /ka-nō-sei/ *potentiality; possibility* [[can + ability = possible; potential] + nature; *-ity* ending], for 4KCWs, the dominant structure is compounding with two 2KCWs combined, such as 自分自身 /ji-bun-ji-shin/ *oneself* [[oneself + one's lot = oneself] + [oneself + someone = oneself]].

1. Introduction

One of the most fundamental characteristics of contemporary written Japanese is its simultaneous employment of multiple scripts, which is

Terry Joyce   ⓘD 0000-0001-9625-1979
School of Global Studies, Tama University, 802 Engyo, Fujisawa, Kanagawa, 252-0805, Japan
E-mail: terry@tama.ac.jp

Hisashi Masuda   ⓘD 0000-0001-8619-6275
Faculty of Health Sciences, Hiroshima Shudo University, 1-1-1 Ozukahigashi, Asaminami-ku, Hiroshima, 731-3195, Japan
E-mail: hmasuda@shudo-u.ac.jp

referred to as 漢字仮名交じり文 /kan-ji-ka-na.ma.jiri.bun/[1] *mixed kanji and kana writing* [kanji + kana + mixed + writing] in Japanese (for fuller accounts of the Japanese writing system (JWS), see Joyce and Masuda, 2018; 2019, as well as Kess and Miyamoto, 1999; Konno, 2013; Smith, 1996; Smith and Schmidt, 1996; Taylor and Taylor, 2014). The four component scripts are morphographic 漢字 /kan-ji/ *kanji* [Han + character], the two separate syllabographic 仮名 /ka-na/ *kana* [provisional + name] scripts of 平仮名 /hira-ga-na/ *hiragana* [smooth + provisional + name] and 片仮名 /kata-ka-na/ *katakana*, [part + provisional + name] and the phonemic alphabet of ローマ字 /rōma.ji/ *Roman letters* [Roman + character], which are supplemented by the small set of Arabic numerals 数字 /sū-ji/ *numbers* [number + character] (Joyce and Masuda, 2018; 2019). Undoubtedly, this unique aspect of the JWS contributes greatly to the highly fungible nature of Japanese written representations (Backhouse, 1984; Joyce, Hodošček, and Nishina, 2012; Joyce and Masuda, 2018; 2019; Miller, 2011; Robertson, 2015; 2017; Smith, 1996; Tranter, 2008). Although Joyce and Masuda (2019) have recently advocated an inclusive notion of intentionality as a promising approach to capturing the diverse motivational factors that influence Japanese graphematic representations, as they equally emphasize, instances of graphematic variation can only be appropriately interpreted with reference to Japanese orthographic conventions. Moreover, as such conventions are closely tied to the historical development of the JWS—from the initial adaption of Chinese characters, the early emergence of the kana scripts, and the relatively recent supplement with rōmaji (Joyce and Masuda, 2018; Lurie, 2012)—there are particularly strong affinities between the scripts and the different lexical strata of the Japanese language (Joyce and Masuda, 2018; 2019; Kageyama and Saito, 2016).

Citing Tamamura (1984) in illustration, Kageyama and Saito (2016) claim that studies of the Japanese language have traditionally distinguished between four 語種 /go-shu/ *word types* [word + type], or lexical strata. They are (1) indigenous 和語 /wa-go/ *Native-Japanese words* (NJ) [Japan + word], (2) 漢語 /kan-go/ *Sino-Japanese words* (SJ) [Han + word], entering from Chinese, (3) 外来語 /gai-rai-go/ *Foreign-Japanese words* (FJ) [outside + come + word], entering from foreign languages since the 16th century, and (4) 混成語 /kon-sei-go/ *hybrid words* [mix + create + word].[2]

---

1. Unless redundant by context, such as within Table listings, Japanese words are represented conventionally and are usually followed by a phonological gloss between slash symbols, / /, English translation in italics, and morpheme meanings and their concatenation within square brackets [ ]. Within the phonological glosses, word boundaries are indicated by spaces, kanji-kanji boundaries by hyphens, and other script boundaries by periods, with macrons, such as ō, indicating long vowels.

2. It should, however, be noted that more recent classifications also cover four types, but the categories differ. Although Shibatani (1990) and Kageyama and Saito

The close affinities between the component scripts and the different lex-ical strata are manifest in a set of general tendencies.[3] Broadly, these are for kanji to represent both SJ and NJ content words as well as NJ verb and adjective stems, for hiragana to represent NJ functional elements such as grammatical markers and inflections, for katakana to represent both FJ and mimetic words, and for rōmaji to represent FJ words and names (Joyce and Masuda, 2019; Kageyama and Saito, 2016).

TABLE 1. Affinities between Japanese lexical strata and JWS component scripts

Stratum	Script	Examples
NJ	Kanji	山 /yama/ *mountain*; 筆 /fude/ *calligraphy brush*
	Kanji-Hiragana	高い /taka.i/ *tall*; 書く /ka.ku/ *to write*
	Hiragana	これ /kore/ *this*; の /no/ *possessive marker*
	Katakana	ワンワン /wanwan/ *doggy*; チカチカ /chikachika/ *flickering, twinkling*
SJ	Kanji	愛 /ai/ *love*; 大学 /dai-gaku/ *university* [big + study]; 正書法 /sei-sho-hō/ *orthography* [correct + write + way]
FJ	Katakana	ミルク /miruku/ *milk*; クラス /kurasu/ *class*; スマートフォン /sumātofon/ *smart phone*
	Rōmaji	PC /pīshī/ *personal computer*; CM /shīemu/ *TV commercial*
Hybrid	Kanji-Kanji	表玄関 /omote-gen-kan/ *front entrance* [NJ+SJ]
	Kanji-Katakana	野菜ジュース /ya-sai.jūsu/ *vegetable juice* [SJ+FJ]
	Hiragana-Katakana	あんパン /an.pan/ *bean-jam bun* [NJ+FJ]

Notes: NJ = native-Japanese; SJ = Sino-Japanese; FJ = foreign-Japanese

---

(2016) continue to recognize the same first three categories (i.e., NJ, SJ and FJ), their fourth category is *mimetic words* that "express non-linguistic sounds or cries or vividly express states or action or physical sensations" (Kageyama and Saito, 2016, p. 12). Usually, they are referred to as 擬音語・擬声語・擬態語 /gi-on-go・gi-sei-go・gi-tai-go/ in Japanese.

3. As one source of deviation from these tendencies, Kageyama and Saito (2016) note that, because the different scripts have distinct perceptual characteristics, such as the stiff and formal impressions of kanji, writers may employ graphematic variants to convey certain nuances. However, as Joyce and Masuda (2019) describe in some detail, there is a wider range of intentionality factors underlying Japanese graphe-matic variation. Accordingly, they treat such script associations as one subcategory of script sensibilities, which is one of their three main factor categories, together with message context and creative representations.

Table 1 presents some examples of these script-lexical strata affinities. Structurally, Table 1 is closely based on Kageyama and Saito's (2016, p. 13) Table 1, entitled 'Classification of word types in traditional Japanese grammar' (as adapted, in turn, from Tamamura, 1984, p. 110), which is primarily from the perspective of the lexical strata. It has, however, been supplemented with a few additional examples from Joyce and Masuda (2019, p. 253) Table 1, entitled 'Examples of standard JWS orthographic conventions', which underscores the same script-strata associations, albeit primarily from the perspective of the JWS's component scripts. While granting that the range of examples in Table 1 may potentially obscure matters, a couple of deeply intertwined points, which are particularly germane to this paper, warrant highlighting. The first point is that, although exact script proportions vary across different genres of written Japanese, kanji are unquestionably the principal component script of the JWS. Indeed, in an interesting study of average script proportions, Igarashi (2007) reports that kanji represented approximately 72%, hiragana 18%, katakana 6% and alphabetic symbols and numbers 4% of the word lists that she extracted from three major newspapers, which, in targeting general adult readerships, closely conform to standard Japanese orthographic conventions.

The second significant point is that, because kanji have deep affinities with the two dominant Japanese lexical strata, both NJ and SJ words, they function as the core units of graphematic representation for a considerable proportion of the Japanese lexicon. Admittedly, this may superficially appear to be merely stating the reason why kanji are the dominant component script within the JWS, but the pluralistic links between kanji and both the NJ and SJ lexical strata are key to understanding the complex nature of Japanese morphographic kanji (Joyce, 2011; Kobayashi, Yamashita, and Kageyama, 2016). Although Kobayashi, Yamashita, and Kageyama (2016, p. 93) tender their remark with specific reference to SJ words, it is essentially impossible to discuss the graphematic representation of the Japanese lexicon as a whole "without some explanation of the *kanji* themselves" (italics in original). It is, therefore, expedient at this point to briefly draw on their succinct account and examples of how kanji became associated with both SJ and NJ words. By their definition, SJ words have entered the Japanese language due to lexical borrowing from the Chinese language; a process that essentially dates back to around the third and fourth centuries to when Chinese characters were initially borrowed and subsequently adapted for written Japanese. Consistent with their morphographic nature in Chinese, kanji represent either a single word or a morpheme, such as 木 meaning *tree*. Also reflecting different historical Chinese pronunciations, this particular kanji is associated with two SJ morphemes or, from the perspective of their phonological values, the two 音読み /on-yo.mi/ *SJ readings* [sound + reading] of /moku/ and /boku/, in different SJ compound words, as in (1).

(1)   /moku/   木馬 /moku-ba/ *wooden horse* [wood + horse]
                材木 /zai-moku/ *timber* [material + tree]
        /boku/   木刀 /boku-tō/ *wooden sword* [tree + sword]
                巨木 /kyo-boku/ *large tree* [giant + tree]

Moreover, as the Japanese language already had NJ words for many of the SJ morphemes represented by kanji, such as the NJ /ki/ for *tree*, it does not require a great leap of imagination to understand how kanji also came to be associated with those NJ morphemes and their phonological values; 訓読み /kun-yo.mi/ *NJ readings* [semantic + reading].[4] As Kobayashi, Yamashita, and Kageyama (ibid.) stress, although some SJ words are monomorphemic and, thus, graphematically represented by a single kanji, such as 茶 /cha/ *tea* and 損 /son/ *loss*, most SJ morphemes are bound morphemes in nature, such that they combine with other SJ morphemes to form compound words.

The Japanese language is particularly interesting from the perspectives of word formation processes and its morphological structures (Kageyama and Saito, 2016; Shibatani, 1990; Tamamura, 1984; 1985). However, as Kageyama and Saito (2016) observe, the application of various word formation processes varies markedly across the different lexical strata. Consistently, although compounding, which Shibatani (1990, p. 237) singles out as being the most productive process by far, is attested with both NJ and SJ elements, it is particularly prominent for SJ words (Kageyama and Saito, 2016; Kobayashi, Yamashita, and Kageyama, 2016). Some sense of the striking differences can be discerned from Joyce, Masuda and Ogawa's (2014) analyses of the graphematic representation codes that they applied to the headwords of the sixth edition of the 広辞苑 /kō-ji-en/ *Kōjien* dictionary (Shinmura, 2008). For example, with C standing for kanji, H for hiragana and K for katakana, 山 was coded as C, 高い as CH, 大学 as 2C, and 山登り /yama-nobo.ri/ *mountain climbing* [mountain + climb] as 2CH. Table 2 shows the ten most frequent graphematic representations codes for the list of Kōjien headwords.[5]

What is particularly striking about these results is that the first three graphematic representation codes of 2C, 3C and 4C (i.e., 2KCWs,

---

4. The official list of characters for general use, known as the 常用漢字表 /jō-yō-kan-ji-hyō/ *Jōyō kanji list* (Agency for Cultural Affairs, 2010), also includes /ko/ as an NJ morpheme in some NJ compound words, such 木陰 /ko-kage/ *shade of tree* [tree + shade]. Kobayashi, Yamashita, and Kageyama (2016, p. 93) refer to it as an "allomorph (apophonic variant)" and acknowledge that "the same character 木 is used in such cases as well".

5. The sixth edition of Kōjien has 232,795 headword entries, but the analyzed list consisted of 215,597 headwords after excluding all kanji that are not on the official jōyō or the Japanese Industrial Standard (JIS) level 1 lists. Of the 1,152 separate graphematic representations codes applied to the list, 578 (50.2%) were unique (i.e., frequency = 1).

Table 2. Ten most frequent graphematic representation codes observed for a list of Kōjien (Shinmura, 2008) headwords (based on Joyce, Masuda, and Ogawa, 2014, p. 188)

Code	Frequency	Percentage	Code	Frequency	Percentage
2C	80,949	37.5	CHCH	4,688	2.2
3C	32,614	15.1	C	4,625	2.1
4C	19,245	8.9	5C	4,495	2.1
2CH	8,916	4.1	CH	4,394	2.0
CHC	5,604	2.6	4K	3,469	1.6

Note: Basic codes are C = kanji, H = hiragana, K = katakana

3KCWs and 4KCWs, respectively) together account for 61.5% of the graphematic representations for the Kōjien headword list. However, although such concatenations of kanji are prototypically characteristic of SJ compounds, it should be stressed immediately that, because their analysis was purely from the perspective of graphematic representation, Joyce, Masuda, and Ogawa (2014) did not seek to explicitly control for lexical strata. Thus, while it is reasonable to assume that the majority of those compound words are SJ compounds, it should also be acknowledged that the frequency counts, particular the 2KCW count, also include some proportion of NJ compound words.

Even though many combinations of two NJ morphemes are graphematically represented with two kanji, pronounced according to their NJ readings, such as 大雨 /ō-ame/ *heavy rain* [big + rain] (Masuda and Joyce, 2018), such combinations more frequently yield graphematic representations that are mixtures of kanji and hiragana.[6] Thus, in contrast to the three most frequent graphematic representations being predominately SJ compounds, the fourth to sixth most frequent codes of 2CH, CHC, and CHCH are likely to be predominately NJ compound words, as illustrated in (2).

(2) 2CH   南向き /minimi-mu.ki/ *facing south* [south + face toward]
             底堅い /soko-gata.i/ *stable (market) after bottoming out*
             [bottom + firm][7]
     CHC   食べ物 /ta.be.mono/ *food* [eat + thing]
             泣き声 /na.ki.goe/ *cry, crying voice* [cry + voice]
     CHCH   立ち読み /ta.chi.yo.mi/ *reading while standing (in store)*
             [stand + read]
             売り買い /u.ri.ka.i/ *trade; buying and selling* [sell + buy]

---

6. This is because the 連用形 /ren-yō-kei/ *infinitive form* [connect + use + form] of many NJ verbs and adjectives consists of a stem and inflection, which are graphematically represented by a kanji and a hiragana, respectively.

7. Kageyama and Saito (2016, p. 20) cite this, together with 高止まり /taka-do.mari/ *remaining high* [high + stop] (2C2H), as evidence of newly coined NJ compounds being common in specialized fields, like the stock market.

In concluding their survey of the word-formation processes and productivity of SJ words, Kobayashi, Yamashita, and Kageyama (2016) single out two reasons why SJ words are so productive (as evidenced in the considerable gaps between the frequencies and percentages of the three most frequent graphematic representation codes compared to the subsequent three codes in Table 2). The first is what Kobayashi, Yamashita, and Kageyama (ibid., p. 129) refer to as a visual factor; namely, "that the meanings of the component morphemes are easily comprehended through the *kanji*" (italics in original). The second reason, which they regard as being the more important, is what they refer to as relaxed restrictions on compound lengths when the compound head is a SJ morpheme. In that context, Kobayashi, Yamashita, and Kageyama (ibid., p. 129) particularly emphasize "the iterative application of compounding rules to produce compounds four or more characters in length and the vigor of affixes that can attach to bases of three or more characters".[8]

Unquestionably, the morphology of Japanese compound words is an especially interesting topic from the perspectives of both writing systems research and the related areas of psycholinguistic research into visual word recognition and the mental lexicon. In light of growing research interest into the representation and retrieval of morphological information within the mental lexicon, Kobayashi et al.'s (2016, p. 129) claim that "kanji play an important role in providing the readers of written Japanese with a visual aid for capturing the meaning of a word at a glance" undoubtedly warrants further empirical investigation. One potentially fertile approach in that respect could be to conduct visual word recognition experiments that utilize the constituent priming paradigm with Japanese compound words of various lengths (Joyce, 2002; Masuda and Joyce, 2018). Moreover, given that kanji are associated with both NJ and SJ morphemes, analyses of the morphological structures of Japanese compound words can potentially further illuminate the intricate nature of morphography in the case of the JWS; a topic of potentially profound significance for writing systems research.

Against such background considerations, this paper reports on the construction of two new databases of 3KCWs and 4KCWs, which have been compiled as components of a larger database project concerned

---

8. However, in order to more appropriately contextualize this comment, it should also be noted that Kobayashi et al.'s (2016) chapter outline only includes sections up to four-character SJ words. As they explain, although it is theoretically possible to construct SJ words of unlimited lengths, such words are inevitably combinations of compound word elements. In illustration, Kobayashi, Yamashita, and Kageyama (2016, pp. 114−115) analyze 新社屋建設案発表会 /shin-sha-oku ken-setsu-an hap-pyō-kai/ *presentation of plan for construction of new company building* according to its component structure, working from its head of 発表会 [[disclose + diagram = presentation] + gathering] for the announcement of the 建設案 [[build + establish = construction] + plan] for the 新社屋 [new + [company + building]].

with Japanese lexical properties (Joyce, Hodošček, and Masuda, 2017; Joyce, Masuda, and Ogawa, 2014). The overarching objective of the larger database project is to compile a database of scale, which can support linguistic and psycholinguistic research on the Japanese lexicon, such as facilitating the selection of stimuli for psycholinguistic surveys and priming experiments (Masuda and Joyce, 2018). Consistent with common practice (Kobayashi, Yamashita, and Kageyama, 2016), the component databases for the larger database project focus on different aspects of the Japanese lexicon, as such compound words according to their overall lengths and targeted lexical properties. For instance, Masuda and Joyce (2005) supplemented a list of 2KCW headwords extracted from the fifth edition of Kōjien (Shinmura, 1995) with various data relating to morphological family sizes, morphological structures and semantic categories, while Masuda, Joyce, et al. (2014) focused on semantic transparency ratings for 2KCWs. The present paper focuses primarily on the analyses of the new database components in terms of the morphological structures of the 3KCWs and 4KCWs, respectively. As the target compound words were extracted according to their graphematic representations, without explicitly controlling for lexical strata, while the majorities of the 3KCWs and 4KCWs will be SJ, inevitably, some proportion of both databases will be either NJ or hybrid compound words. After briefly outlining the extracting and cleaning of the two database lists in Section 2, Sections 3 and 4 present the results of analyzing the morphological structures of the 3KCW and 4KCWs, respectively. The paper ends with a short section of concluding remarks.

2. List Extraction and Cleaning

Although the analyzed lists of 3KCWs and 4KCWs were extracted on separate occasions, the two-stage extraction procedures were identical in both cases. During the respective first stages, all the relevant compound words were extracted from the set of corpus word lists (CWLs) that Joyce, Hodošček, and Nishina (2012) compiled from the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Joyce, Hodošček, and Masuda, 2017; Maekawa et al., 2013). Joyce et al.'s (2012) CWLs are grouped according to both the two word-units definitions[9] and the word class divisions used within the BCCWJ project, and all CWL files,

---

9. The main lexical demarcation employed with the BCCWJ is a somewhat elusive one in distinguishing between short-unit words (SUWs) and long-unit words (LUWs). Although the short-long labels evoke a length-based contrast, as Joyce, Masuda, and Ogawa (2014) explain, the distinction is essentially of lexical status, such that SUWs include both bound morphemes and simple words (dictionary headwords) and LUWs are complex words and phrases.

apart from the proper noun files, were examined to check for the possible presence of target compound words. In addition to recording the CWL source file, all of the CWL's lexical information was retained for reference in analyzing the compound words. This information includes columns for the underlying lemma entry, lemma length (used to extract target compounds), number of graphematic variants, etymology code, BCCWJ frequency of the lemma, orthographic base (graphematic variants of a lemma), orthographic base pronunciations, lengths of orthographic bases, BCCWJ frequency of the orthographic base, and ratio of total lemma frequency covered by a particular orthographic base form. Stage 1 processing resulted in spreadsheets of 171,123 rows of 3KCWs and 298,944 rows of 4KCWs.

The substantial disparity in the numbers of spreadsheet rows for the 3KCWs and 4KCWs extracted from the CWLs is consistent with the analyses of graphematic representation codes that Joyce, Hodošček, and Masuda (2017) also conducted for Joyce et al.'s (2012) CWLs. Focusing only on the relevant long-unit word (LUW) data, even though the first and second most frequent graphematic representation codes by types counts were 4C (15.4%) and 3C (9.3%), respectively, by token counts, the 3C code was only the eighth most frequent (3.1%) and the 4C code did not feature amongst the top ten codes at all. Those findings indicate that, although there are far fewer 3KCWs than 4KCWs within the Japanese lexicon overall, 3KCWs generally tend to occur more frequently than 4KCWs.

In order to derive lists of more practical lengths for analyses, the respective second stages commenced by first applying the criterion that the BCCWJ lemma frequencies (token counts) should be either equal to or greater than 10. Moreover, reflecting the automatic nature of the methods used in extracting the CWL source corpus, additional cleaning work was required to remove some non-words, some proper nouns and to merge for cases of lemma replications. Accordingly, Stage 2 processing resulted in database lists of 23,046 3KCW-lemmas and 23,159 4KCW-lemmas. Although the application of the frequency criterion yielded highly comparable lists in terms of the overall numbers of compound word lemmas that each database component contains, naturally, the impact of eliminating compound words with frequencies of less than 10 was far greater in the case of the 4KCWs. That is, although Stage 2 processing for the 3KCWs yielded a list that was 13.5% of the Stage 1 extracted list, Stage 2 processing for the 4KCWs yielded a list that contained only 7.75% of the Stage 1 extracted list. It should also be noted that while the distributions of lemma frequencies are generally consistent for both database lists, with both being typical of corpus frequencies, the 3KCWs are generally of higher token frequency counts compared to the 4KCWs, as the plots of log-transformed frequencies in Figure 1 indicate. More specifically, for the 3KCWs, the frequency range is from 10

to 18,395 with a mean of 88.5 and median of 25, while for the 4KCWs, the frequency range is from 10 to 4,127 with a mean of 43.1 and a median of 20.
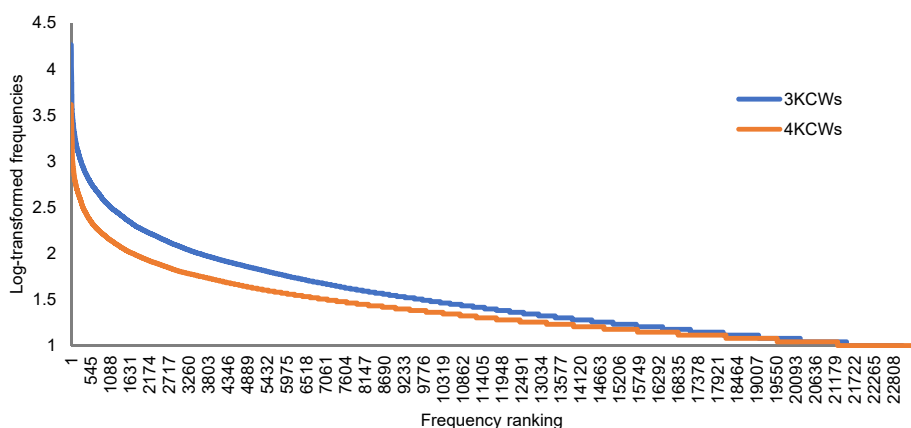
FIGURE 1. Log-transformed lemma frequencies of the 3KCWs and 4KCWs

Even though the BCCWJ's original lexical strata codes (i.e., NJ, SJ or hybrids) are retained within the respective databases compiled from the CWLs, the primary criterion for inclusion has been the appropriate lemma length of 3KCWs and 4KCWs, respectively. Thus, while acknowledging that both database lists contain some proportions of NJ and hybrid compound words and that awareness concerning the lexical stratum of the components often greatly informs the appropriate classifications of the compound words, the analyses of morphological structures reported in the subsequent sections do not explicitly consider the lexical stratum of the component elements. The conducted analyses of both database lists adopted similar conventions for denoting the constituent component kanji, which were designated as A, B, and C (3KCWs), as well as D (4KCWs), respectively, with square-brackets used to indicate internal structures, such as [AB]+C to indicate a 2KWC with a C addition and [AB]+[CD] to indicate a combination of two 2KCWs. Moreover, as Kobayashi, Yamashita, and Kageyama (2016, p. 108) emphasize, with SJ morphemes, in particular, it can often be quite difficult to discern both a morpheme's status, as either a free word or bound element, and the word-formation process that underlies a particular compound word, as either involving compounding or affix-derivation. Accordingly, in considering the appropriate classification of all compound words, we have also checked for alternative structures. To that end, all compound words were initially segmented and the component kanji

were subsequently recombined to consider for all possible structures. For example, as 農業 /nō-gyō/ *agriculture* [agriculture + business] and 業者 /gyō-sha/ *trader, business person* [business + person] both exist as 2KCWs, it is necessary to consider all component meanings and usage patterns to determine that [AB]+C is the more coherent interpretation of 農業者 /nō-gyō-sha/ *agricultural worker* [agriculture + business + person].[10]

3.   The Morphological Structures of the 3KCW Database

Although Joyce and Masuda (2019) tendered an initial report about compiling this database list of 23,046 3KCWs, this paper describes the results of analyzing their morphological structures in a little more detail. In addition to presenting a summary table of the morphological structures, Section 3.1 includes a table of the top 20 most frequent 3KCWs, as well as some general analyses of the A and C additions to 2KCWs. Three further sub-sections focus on the morphological structures, with Section 3.2 on the primary structure of [AB]+C, Section 3.3 on the secondary structure of A+[BC], and Section 3.4 on the remaining 3KCW structures.

3.1.   Morphological Structures of the 3KCW Database: Summary and A + C Additions

Table 3 presents the breakdown of the 3KCW database list according to their morphological structures, with both type counts and their corresponding percentages. As the morphological structures of 3KCWs are generally transparent, it has been possible to confidently classify the database list according to eight morphological structures.[11] As Table 3 clearly indicates, the primary morphological structure of [AB]+C is highly dominant in accounting for 77.1% of the database list. In contrast, the secondary structure of A+[BC] only accounts for 21.3% overall, which is about one-third of the primary structure's percentage. However, taking the primary and secondary structures together, they account for the vast majority of 3KCWs, at 98.4% for the database list, with six other structures underlying the remaining 1.6%. Firmly underscoring the profound significance of 2KCWs within the Japanese lexicon

---

10. Although we regard the [AB]+C classification as being the more plausible interpretation, we are also planning to conduct psycholinguistic surveys to investigate the extent to which alternative structures might be activated in the processing of such compound words.

11. Table 3 also includes an adjustment category of multiple types for a few 3KCWs that are open to alternative analyses.

(Joyce, 2011; Joyce, Hodošček, and Masuda, 2017; Kobayashi, Yamashita, and Kageyama, 2016; Nomura, 1975; 1988), the majority of 3KCWs are 2KCWs combined with an additional morpheme; either predominately attached to the end or, in considerable cases, inserted at the beginning.

Table 3. Breakdown of the morphological structures in the 3KCW database

Morphological structure	Type counts	Percentage
[AB]+C	17,761	77.1
A+[BC]	4,904	21.3
[A(C*)]+[BC] (with (C*) omitted)	154	0.7
Non-divisible	93	0.4
Phonological transcription (当て字)	64	0.3
Monomorphemic (熟字訓)	45	0.2
A+B+C	25	0.1
[AB]+[(A*)C] (with (A*) omitted)	15	0.1
Multiple types (count adjustment)	-15	−0.1
Total	**23,046**	**100**

Table 4 presents the 20 most frequent 3KCWs based on token frequency counts, which indicates that frequency is independent of morphological structure. Although the primary morphological structure of [AB]+C is the most frequent among these most frequent 3KCWs, which is consistent with the overall analysis results, other morphological structures are also associated with highly frequent 3KCWs, such as 雰囲気 /fun-i-ki/ *mood; ambience* [atmosphere + surround + spirit], which is classified as non-divisible. Although each of the SJ morphemes contributes semantically to some degree to the overall meaning of this 3KCW, its original etymology is no longer obvious.

Understandably, a sizeable proportion, at 12.0% of the 3KCWs are combinations of number kanji with various numerical units and classifiers and, as may be also discerned from Table 4, some of these are of high frequencies, such as 三十分 / san-jip-pun/ *thirty minutes* [[three + ten = thirty] + minutes] and 十二月 /jū-ni-gatsu/ *December; 12 months* [[ten + two = twelve] + month].

Before turning to the dominant primary and secondary morphological structures, as the vast majority of 3KCWs involve either a single SJ or NJ morpheme being added to an existing 2KCW, it is beneficial to also note Masuda and Joyce's (2019) separate analyses of the A and C additions. Notwithstanding the challenges, with most kanji being associated with both multiple SJ and multiple NJ morphemes and that the status of

TABLE 4. 20 most frequent 3KCWs by token frequency counts for the orthographic base

3KCW	Structure	Gloss	Translation and explanation	Frequency
大丈夫	A+[BC]	dai-jō-bu	problem-free [big + [stature + man = healthy]]	16,861
可能性	[AB]+C	ka-nō-sei	possibility [[can + able = possible] + -ity ending]	13,555
不思議	A+[BC]	fu-shi-gi	mysterious [negative + [think + debate = conjecture; guess]]	13,044
雰囲気	Non-divisible	fun-i-ki	mood; ambience [atmosphere + surround + spirit]	11,427
三十分	[AB]+C	san-jip-pun	thirty minutes [[three + ten = thirty] + minutes]	11,042
十二月	[AB]+C	jū-ni-gatsu	December; 12 months [[ten + two = twelve] + month]	10,325
十一月	[AB]+C	jū-ichi-gatsu	November; 11 months [[ten + one = eleven] + month]	9,739
具体的	[AB]+C	gu-tai-teki	concrete, specific [[means + substance = tangible] + -ic adjectival noun (AN) ending]	9,334
基本的	[AB]+C	ki-hon-teki	fundamental, basic [[foundation + base = basics] + -ic AN ending]	8,251
第一項	[AB]+C	dai-ik-kō	first item [[ordinal number + one = first] + item; clause]	7,261
積極的	[AB]+C	sek-kyoku-teki	positive; active [[amass + poles = active, positive] + -ic AN ending]	7,060
大統領	A+[BC]	dai-tō-ryō	president [big + [govern + territory = ruler; leader; consul]]	6,992
出来事	[AB]+C	de-ki-goto	incident; event [[go out + come = occurrence; happening] + thing]	6,189
一般的	[AB]+C	ip-pan-teki	general, typical [[one + general = general; ordinary] + -ic AN ending]	5,932
不可欠	Non-divisible	fu-ka-ketsu	indispensable; essential [negative + can + lack; fail]	4,209
十五日	[AB]+C	jū-go-nichi	15th; 15 days [[ten + five] + day]	3,967
高齢者	[AB]+C	kō-rei-sha	elderly person/people [[high + age = old] + person]	3,902
青少年	[A(C*)]+[BC]	sei-shō-nen	youths, young people [[green + years* = youths] + [few + years = youth]]	3,751
二十日	[AB]+C	hatsu-ka	20th; 20 days [[two + ten] + day]	3,608
小学校	A+[BC]	shō-gak-kō	elementary school [small + [study + school = school]]	3,540

any given kanji can vary across different 3KCWs,[12] Masuda and Joyce analyzed the additional A and C components according to their morpheme status, as either free, bound or affix morphemes. The analysis results are presented in Table 5.

TABLE 5. Results of analysing A and C additional components in terms of their morpheme status

Morpheme status	A additions		C additions	
	Type count	Percentage	Type count	Percentage
Free	360	55.0	369	44.0
Bound	225	34.4	401	47.9
Affix	70	10.7	68	8.1
Total	**655**	**100.0**	**838**	**100.0**

3.2.   Primary Morphological Structure of [AB]+C 3KCWs

As Table 3 vividly attests, 3KCWs overwhelmingly conform to the morphological structure of [AB]+C, where a single morpheme is appended to an existing 2KCW. Accordingly, Table 6 first presents the ten most frequent C-additions in terms of their type counts, which indicates their productivity in combining with multiple 2KCWs, then Table 7 presents the most frequent 3KCWs by token counts, for each of the 10 most frequent C-additions.

Being wholly consistent with Kobayashi et al.'s (2016, p. 127) comment that 的 /teki/ *AN ending*[13] is "a representative, highly productive Sino-Japanese affix that combines with a variety of bases," it is not in the least surprising to find that it is the most productive of the C-additions observed within the 3KCW database list. Indicative of its wide applicability, 的 is a C-component of 3KCWs across the database's entire frequency range. In addition to being a C-addition to four of the 20 most

---

12. Kobayashi, Yamashita, and Kageyama (2016, pp. 95–96) classify one-character SJ morphemes as free (会 /kai/ *meeting*) or bound—either connectives (運転中 /unten-chū *while driving*), or the bases of verbs (信じる /shin.jiru/ *believe*), of ANs (急な /kyū.na/ *abrupt*), of adverbs (実に /jitsu.ni/ *actually*), of adnominal/adverbial modifiers (単なる /tan.naru/ *mere*). Kobayashi et al. also regard some bound morphemes as affixes due to their positional constraints, such as 最 /sai/ *most* as a prefix of 最先端 /sai-sen-tan/ *cutting edge* [most + front + edge] (p. 108).

13. As Kageyama and Saito (2016, p. 18) note, the lexical category of adjectival noun does not exist in English or other European languages. While morphologically a noun, it can function syntactically as an adjective with -だ /-da/ in predicates and -な /-na/ adnominally.

TABLE 6. Top ten most frequent C-additions to [AB]+C 3KCWs by type counts

C-addition	Meaning	Type count
的	*-ic* AN ending	873
者	*-er* person-indicating ending	685
等	etc.; and so forth	577
性	*-ity* ending; nature	498
中	in/during [place or time]	352
化	*-ization* verbal noun (VN) ending	294
後	after	253
達	pluralizing ending	244
上	above; in terms of	239
人	*-er* person-indicating ending	227

TABLE 7. Most frequent [AB]+C 3KCWs by token counts, for each of the most frequent C-additions

3KCW	Gloss	Translation and explanation	Frequency
具体的	gu-tai-teki	concrete [[means + substance] + AN ending]	9,334
高齢者	kō-rei-sha	elderly person/people [[high + age] + person]	3,902
整備等	sei-bi-tō	maintenance etc. [[organize + equip] + etc]	608
可能性	ka-nō-sei	possibility [[can + able] + *-ity* ending]	13,555
世界中	se-kai-jū	around the world [[world + world] + thoughout]	2,034
生活化	sei-katsu-ka	living [[life + active] + VN ending]	1,195
十年後	jū-nen-go	after 10 years; 10 years later [[ten + year] + after]	476
子供達	ko-domo-tachi	children [[child + accompany] + pluralizer]	886
事実上	ji-jitsu-jō	as a matter of fact [[thing + real] + in terms of]	1,396
外国人	gai-koku-jin	foreigner [[outside + country] + *-er* person]	2,361

frequent 3KCWs (Table 4), some other 3KCW examples that vary in terms of their token frequencies are listed in (3).

(3) 比較的 /hi-kaku-teki/ *comparatively*   3,515
      [[compare + contrast] + *-ic* AN]
      国際的 /koku-sai-teki/ *international*   2,106
      [[country + occasion; side] + *-ic* AN]

本質的	/hon-shitsu-teki/ *intrinsic; substantial*	1,026
	[[true + quality] + *-ic* AN]	
潜在的	/sen-zai-teki/ *implicit, latent*	506
	[[conceal + exist] + *-ic* AN]	
挑戦的	/chō-sen-teki/ *challenging; provocative*	99
	[[contend + battle] + *-ic* AN]	

As Kobayashi, Yamashita, and Kageyama (2016) point out, 的 combines with various bases, as their examples in (4) illustrate, and, consistently, it is one of the most frequent D-additions to 4KCWs.

(4) 私的   /shi-teki/ or /watashi-teki/ private; personal [I + -ic AN]
     活動的   /katsu-dō-teki/ *active; dynamic* [[active + move] + *-ic* AN]
     政治家的   /sei-ji-ka-teki/ *politician-like*
                 [[[politics + rule] + person] + *-ic* AN]
     共産主義的 /kyō-san-shu-gi-teki/ *communistic*
                 [[[together + produce] + [principle + meaning]] + *-ic* AN]
     草分け的   /kusa-wa.ke.teki/ *pioneering* [[grass + divide] + *-ic* AN]
     カリスマ的 /karisuma.teki/ *charismatic* [charisma + *-ic* AN]

As Table 6 shows, the tenth most productive C-addition is 人 person, reflecting its generic sense, but as Kobayashi, Yamashita, and Kageyama (ibid., p. 127) point out, it is associated with two SJ morphemes. The first is /jin/, which attaches to both nouns and stems of ANs, such as the examples in (5).

(5) 外国人 gai-koku-jin *foreigner* [[outside + country] + -er person]
     芸能人 gei-nō-jin *performer* [[perform + talent] + *-er* person]
     有名人 yū-mei-jin *famous person* [[possess + name] + *-er* person

The second SJ morpheme is /nin/, which only attaches to verbal nouns (VN),[14] such as the examples in (6). A further restriction is that while /nin/ attaches to NJ bases, such as 受け取り人 /u.ke.to.ri.nin/ *recipient* [[receive + take] + person], /jin/ does not, apart from the single exception of 暇人 /hima-jin/ *person of leisure* [leisure + person].

(6) 通行人 /tsū-kō-nin/ *passerby* [[pass through + go] + person]
     弁護人 /ben-go-nin/ *advocate; defender*
              [[speech + safeguard] + person]
     管理人 /kan-ri-nin/ *manager; administrator*
              [[control + arrange] + person]

---

14. As Kageyama and Saito (2016, p. 18) also stress, the verbal noun (VN) is another lexical categories that does not exist in European languages. Kagayama and Saito describe VNs as "hybrid category" of a noun that can function as a verb when combined with the dummy verb する /suru/.

3.3.   Secondary Morphological Structure of A+[BC] 3KCWs

Although not as common as the primary structure of [AB]+C 3KCWs, the secondary structure of A+[BC] accounts for approximately one-fifth (21.3%) of the 3KCW database list. Table 8 presents the ten most frequent A-additions in terms of their type counts, while Table 9 presents, for each of the ten most frequent A-additions, the most frequent A+[BC] 3KCWs with the respective A-additions.

TABLE 8. Top ten most frequent A-additions to A+[BC] 3KCWs by type counts

A-addition	Meaning	Type count
御	honorific prefix	430
大	large; big	313
各	each; every	152
不	negative prefix *non-*	143
新	new	127
一	one	126
無	negative prefix *un-*, *non-*	95
同	same	93
諸	various; several	90
全	all; whole	86

TABLE 9. Most frequent A+[BC] 3KCWs by token counts, for each of the most frequent A-additions

3KCW	Gloss	Translation and explanation	Frequency
御指摘	go-shi-teki	as you indicate [honorific + [point + pinch]]	2,171
大丈夫	dai-jō-bu	problem-free [big + [stature + man = healthy]]	16,861
各地域	kaku-chi-iki	each region [each + [ground + region]]	388
不思議	fu-shi-gi	mysterious [negative + [think + debate]]	13,044
新幹線	shin-kan-sen	bullet train [new + [trunk + line]]	1,118
一時間	ichi-ji-kan	one hour [one + [time + interval]]	6,515
無意識	mu-i-shiki	unconsciousness [un- + [mind + know]]	1,263
同級生	dō-kyū-sei	classmate [same + [rank; class + student]	840
諸外国	sho-gai-koku	various foreign countries [various + [out + country]	744
全世界	zen-se-kai	whole world [all + [[world + world]]	619

While Kobayashi, Yamashita, and Kageyama (2016, p. 123) acknowledge that there are considerable cases of SJ words "where the distinction between affix and compound constituent is not clear," they also stress that A-additions often represent substantive semantic concepts. Indeed, they provide a number of examples, which they organize according to five semantic functions, including (a) limiting or modifying the base meaning, (b) verbal meaning that corresponds to the base noun's argument, (c) limiting the base noun's reference, (d) adverbially modifying a predicate-like base, and (e) indicating negation. Examples for each of these five semantic functions are given in (7).

(7)   好成績   /kō-sei-seki/ *good results* [pleasing + [become + achievements]]
      反体制   /han-tai-sei/ *anti-establishment* [opposite + [body + system]]
      本製品   /hon-sei-hin/ *this product* [this; main + [manufacture + goods]]
      急成長   /kyū-sei-chō/ *rapid growth* [rapid + [become + long]]
      未経験   /mi-kei-ken/ *inexperienced* [not yet + [pass thru + effect]]

Instructively, Kobayashi, Yamashita, and Kageyama (ibid., pp. 126–127) differentiate between the four A-additions that signify negative senses in terms of their nuances and the categories of bases to which they attach. Consistent with its fourth place ranking amongst the most productive A-additions, Kobayashi, Yamashita, and Kageyama (ibid.) comment that 不 /fu/ and /bu/ *negative* is the most productive and attaches to nouns, adjectival nouns and verbal nouns, as in (8), respectively.

(8)   不景気   /fu-kei-ki/ *recession* [negation + [view + spirit; atmosphere]]
      不確実   /fu-kaku-jitsu/ *uncertain* [negation + [confirm + reality]]
      不承知   /fu-shō-chi/ *disapproval* [negation + [acquiesce + know]]

The next most productive A-addition with negative connotations is 無 /mu/ *lacking, non-existent*, which attaches to nouns and verbal nouns, but not adjectival nouns, as in (9).

(9)   無関心   /mu-kan-shin/ *unconcerned* [lacking + [connection + heart]]
      無関係   /mu-kan-kei/ *unrelated* [lacking + [connection + connection]]

As an A-addition of 45 3KCWs within the database list, the third most productive of the A-additions with negative connotations is 未 /mi/ *not yet*, which attaches to nouns and verbal nouns, but not adjectival nouns, as in (10).

(10) 未成年   /mi-sei-nen/ *not of age* [not yet [become + age]]
      未解決   /mi-kai-ketsu/ *unresolved*
              [not yet + [unravel; solve + decide; fix]]

While only attested as an A-addition to 38 A+[BC] 3KCWs within the database list, 非 /hi/ *negation* also attaches to nouns, adjectival nouns and verbal nouns, as in (11) respectively.

(11)　非人情　/hi-nin-jō/ *inhuman* [negation + [person + feelings]]
　　　非合法　/hi-gō-hō/ *unlawful* [negation + [fit; suit + law, rule]
　　　非公認　/hi-kō-nin/ *unauthorized*
　　　　　　　[negation + [public; official + acknowledge]]

3.4.   Other 3KCW Morphological Structures

Although our analysis of the morphological structures of the 3KCW database reveals that the two structures of [AB]+C and A+[BC] account for the vast majority (98.4%) of 3KCWs, as Table 3 also indicates, six other morphological structures underlie a small percentage of 3KCWs. Accordingly, this section turns to present examples of 3KCWs that conform to those other morphological structures.

　　Albeit on a distinctly smaller scale (0.7%), the third most frequent morphological structure is [A(C*)]+[BC], where the C-component of an [AC] 2KCW is omitted and the resultant A is attached to a related [BC] 2KCW. The practice of omitting the C-component of an [AC] 2KCWs is undoubtedly a form of clipping that is common with SJ words (Kobayashi, Yamashita, and Kageyama, 2016, p. 128).[15] As such, superficially, this structure may appear to resemble the A+[BC] structure, in the sense, that it effectively involves an A-component being inserted before a [BC] 2KCW. It is, however, appropriate to differentiate them, because the [A(C*)]+[BC] structure crucially hinges on the semantic relationship between the [AC] and [BC] 2KCWs, due to their shared C-component, as the examples in (12) illustrate.

(12)　[A(C*)]+[BC] (with (C*) omitted)
　　　視聴覚　/shi-chō-kaku/ *audiovisual* [視覚 vision + 聴覚 hearing]
　　　入出国　/nyū-shutsu-koku/ *immigration*
　　　　　　　[入国 enter country + 出国 depart county]

---

15. Clipping with SJ words most typically involves 4KCWs being shortened to 2KCWs, such as 模擬試験 /mo-gi-shi-ken/ *practice test* → 模試 /mo-shi/ (A + C) or 高等学校 /kō-tō-gak-kō/ *high school* → 高校 /kō-kō/ (A + D). One important consequence of such clipping processes is that the resultant 2KCWs tend to have far higher frequencies than the corresponding 4KCW, such as 就活 /shū-katsu/ *job hunting* [take position + activity] which is derived by clipping from 就職活動 /shū-shoku-katsu-dō/ *job hunting* [position + post + lively + move].

The fourth category of non-divisible is necessary to handle the small set of exceptions (0.4%). Some 3KCWs are classified as non-divisible, because the compound word's etymology and morphological structure are not clear, even though the meanings of the component morphemes are usually related to the overall meaning. Other 3KCWs classified as non-divisible are the results of clipping processes applied to longer compound words. One example of each is presented in (13).

(13)   Non-divisible
        方程式   /hō-tei-shiki/ *equation; formula*
               [direction + formula + expression]
        食洗機   /shoku-sen-ki/ *dishwasher* ← 食器洗浄機
               [[eat + ware = dishes] + [[wash + clean = washing] + machine]]

The fifth category is phonological transcriptions (0.3%), known as 当て字 /a.te.ji/ *phonological transcription* [apply + character] in Japanese, which refers to the convention of phonologically representing a word's syllables with kanji. Although phonological transcriptions are essentially a form of the rebus principle, the individual kanji used for such graphematic representations often have some degree of semantic relevance to the word's meaning, such as in the first example in (14), but sometimes less so, as in the second example.

(14)   Phonological transcriptions (当て字)
        歌舞伎   /kabuki/ *kabuki; Japanese classical drama* [sing + dance + art]
        目論見   /mokuromi/ *plan; scheme; plot* [eye + argument + see]

The sixth category is monomorphemic words (0.2%), known as 熟字訓 /juku-ji-kun/ *monomorphemic word* [compound + character + semantic translation] in Japanese, which refers to the convention of representing the meaning of an NJ word with kanji that are semantically related. In contrast to phonological transcriptions where the kanji are representing the syllables of the word, there is usually no phonological correspondence between the elements of the graphematic representation, but the meanings of the component kanji are related to the word's meaning. The first example in (15) may be regarded as the prototypical example that is frequently cited in illustration.

(15)   Monomorphemic words (熟字訓)
        五月雨   /samidare/ *early summer rain* [five + month + rain]
        波止場   /hatoba/ *wharf; quay* [wave + stop + place]

The seventh morphological structure is A+B+C (0.1%), as the concatenation of three morphemes that together constitute some form of set or may be regarded as exemplars of the compound word's meaning, as both the examples in (16) indicate.

(16)   A+B+C
      衣食住   /i-shoku-jū/ *necessities of life* [clothing + food + shelter]
      産官学   /san-kan-gaku/ *industry, government and academia*
            [industry + government + academia]

The eighth and final morphological structure is [AB]+[(A*)C] (0.1%), where the A-component of an [AC] 2KCW is omitted and the resultant C is attached to an [AB] 2KCW. Like the [A(C*)]+[BC] structure, the omitting of the A-component of an [AC] 2KCW is also a form of clipping. Also similar to the [A(C*)]+[BC] structure, it is appropriate to differentiate this from the primary morphological structure of [AB]+C 2KCWs, because the [AB]+[(A*)C] structure also hinges on the semantic relationships between the [AB] and [AC] 2KCWs, due to their shared A-component, as the examples in (17) illustrate.

(17)   [AB]+[(A*)C] (with (A*) omitted)
      国内外   /koku-nai-gai/ *domestic + foreign* [国内 domestic + 国外 foreign]
      十五六   /jū-go-roku/ *15 or 16* [十五 15 + 十六 16]

4.   Morphological Structure Results for the 4KCW Database

Having presented the results of analyzing the morphological structures of the 3KCW database component in some detail, this paper now turns to present the results for the 4KCW database. Adopting a similar organization to the previous section, Section 4.1 starts with a summary table of the morphological structures and a table of the top 20 most frequent 4KCWs. Four further sub-sections focus on the various morphological structures, with Section 4.2 on the primary structure of [AB]+[CD], Section 4.3 on the second structure of [ABC]+D, Section 4.4 on the tertiary structure of A+[BCD], and Section 4.5 on the remaining 4KCW structures.

4.1.   Morphological Structures of the 4KCW Database: Summary

Table 10 presents the breakdown of the database list of 23,159 4KCWs according to their morphological structures, with both type counts and their corresponding percentages.

As the morphological structures of 4KCWs are also generally highly transparent, it has been possible to confidently classify the database

TABLE 10. Breakdown of the morphological structures in the 4KCW database

Morphological structure	Type counts	Percentage
[AB]+[CD]	19,805	85.3
[ABC]+D	2,809	12.1
A+[BCD]	449	1.9
Non-divisible	23	0.1
[A(CD*)]+[BCD] (with (*CD) omitted)	18	0.1
[A(D*)]+[B(D*)]+[CD] (with both (*D) omitted)	16	0.1
A+B+C+D	16	0.1
Phonological transcriptions (当て字)	14	0.1
[AB]+C+D	6	0.0
Monomorphemic (熟字訓)	2	0.0
[A(D*)]+[BCD] (with (*D) omitted)	1	0.0
Total	**23,159**	**100**

list according to 11 morphological structures.[16] Similar to the results for the 3KCWs, the analyses of the morphological structures within the 4KCWs reveals that one structure dominates in accounting for 85.3% of the 4KCW types. However, in the case of 4KCWs, the primary morphological structure is [AB]+[CD], which is consistent with Kobayashi et al.'s (2016, p. 113) comment that "four-character S-J words are words composed of four S-J morphemes, which are typically divided into two words, each consisting of two morphemes". This primary structure also attests to the immense significance of 2KCWs within the Japanese lexicon (Joyce, 2011; Joyce, Hodošček, and Masuda, 2017; Kobayashi, Yamashita, and Kageyama, 2016; Nomura, 1975; 1988).

Reflecting the even greater dominance of the [AB]+[CD] structure, in contrast, the secondary structure of [ABC]+D and the tertiary structure of A+[BCD] account for 12.1% and 1.9%, respectively, of all 4KCW structures. Naturally, there are parallels between these morphological structures and the primary and secondary structures of 3KCWs, as they also involve combining an additional morpheme with an existing compound word and the marked preference is for attaching that additional morpheme to the end rather than inserting at the beginning. However, reflecting the even greater dominance of the primary structure, the secondary and tertiary structures are relatively less common for 4KCWs.

---

16. Claiming that the structures of 4KCWs "can be categorized by the patterns of binary branching structures," Kobayashi, Yamashita, and Kageyama (2016, pp. 114–115) list nine patterns under four types, following Nomura (1975). Reflecting its importance, the first type is [AB]+[CB], the second is of 3KCWs plus additions (i.e., [ABC]+D and A+[BCD]), the third involves combinations (i.e., [ACD*]+[BCD] and [AD*]+[BD*]+[CD]), and the fourth is A+B+C+D. However, their list does not include either phonological transcriptions or monomorphemic words.

TABLE 11. 20 most frequent 4KCWs by token frequency counts

4KCW	Structure	Gloss	Translation and explanation	Frequency
自分自身	[AB]+[CD]	ji-bun-ji-shin	oneself [[oneself + one's lot = oneself] + [oneself + someone = oneself]]	3,288
三十一日	[ABC]+D	san-jū-ichi-nichi	31st; 31 days [[three + ten + one] + day]	3,238
二十五日	[ABC]+D	ni-jū-go-nichi	25th; 25 days [[two + ten + five] + day]	3,058
二十一日	[ABC]+D	ni-jū-ichi-nichi	21st; 21 days [[two + ten + one] + day]	2,718
二十四日	[ABC]+D	ni-jū-yok-ka	24st; 24 days [[two + ten + four] + day]	2,716
二十三日	[ABC]+D	ni-jū-san-nichi	23st; 23 days [[two + ten + three] + day]	2,660
二十二日	[ABC]+D	ni-jū-ni-nichi	22rd; 22 days [[two + ten + two] + day]	2,659
二十八日	[ABC]+D	ni-jū-hachi-nichi	28th; 28 days [[two + ten + eight] + day]	2,642
都道府県	A+B+C+D	to-dō-fu-ken	administrative divisions [Tokyo + Hokkaido + Osaka/Kyoto + prefectures]	2,621
二十六日	[ABC]+D	ni-jū-roku-nichi	26th; 26 days [[two + ten + six] + day]	2,587
二十七日	[ABC]+D	ni-jū-shichi-nichi	27th; 27 days [[two + ten + seven] + day]	2,495
二十九日	[ABC]+D	ni-jū-ku-nichi	29th; 29 days [[two + ten + nine] + day]	2,492
政府委員	[AB]+[CD]	sei-fu-i-in	ministerial aide [[politics + government] + [committee + member]]	2,336
中小企業	[AB]+[CD]	chū-shō-ki-gyō	small-medium companies [[middle + small] + [plan + business = company]]	2,176
二千一年	[ABC]+D	ni-sen-ichi-nen	2001 [[two + thousand + one] + year]	2,053
金融機関	[AB]+[CD]	kin-yū-ki-kan	financial institutions [[money + melt] + [mechanism + concern]]	2,051
携帯電話	[AB]+[CD]	kei-tai-den-wa	mobile phones [[carry + belt = mobile] + [electric + talk = phone]]	2,025
人間関係	[AB]+[CD]	nin-gen-kan-kei	human relations [[person + interval = human] + [connect + connect]]	1,907
二千二年	[ABC]+D	ni-sen-ni-nen	2002 [[two + thousand + two] + year]	1,861
一生懸命	[AB]+[CD]	is-shō-ken-mei	with utmost effort [[one + life = lifetime] + [depend + fate = effort]]	1,861

601

Table 11 presents the 20 most frequent 4KCWs by token frequency counts. Comparing Table 11 with Table 3, which present the 20 most frequent 3KCWs, might initially appear to somewhat undermine the claim advanced earlier that morphological structures are independent of word frequencies. Even though Table 10 clearly indicates that the [AB]+[CD] structure is the highly dominant one for 4KCWs, at 85.3% of all types, 12 of the top 20 4KCWs have [ABC]+D structures and only seven have [AB]+[CD] structures. However, it should be noted that ten of those [ABC]+D 4KCWs are referring to dates of the month, such as 三十一日 /san-jū-ichi-nichi/ *the thirty-first; 31 days*, where the ABC kanji represents 31 and the D-addition represents day, with the other two [ABC]+D 4KCWs being year designations (e.g., 二千一年 2001). As such, their occurrences within the top 20 4KCWs should be attributed to the relative frequency levels of these compound word lemmas within the BCCWJ-based CWLs that are the basis for the 3KWC and 4KCW database lists. That is, while these particular 4KCWs are of high frequencies among the 4KCW database list, as noted earlier, the 4KCWs are generally of lower frequencies compared to the 3KCWs. It is also germane in this context to note that, although 12.0% of the 3KCWs are combinations of number kanji with various numerical units and classifiers, only 1,134 (4.9%) of the 4KCW list are of such combinations, with another 332 (1.4%) 4KCW that are only numbers. Of the 1,134 4KCWs that are combinations of a number and a unit or classifier, understandably, 991 (87.4%) of those are [ABC]+D structures.[17]

4.2. Primary Morphological Structure of [AB]+[CD] 4KCWs

As the summary results in Table 10 incontestably indicate, the primary morphological structure of 4KCWs is [AB]+[CD], where two 2KCWs are combined into a larger compound unit. Notwithstanding Kobayashi et al.'s (2016, p. 117) observation that the semantic head of most [AB]+[CD] 4KCWs is on the right-side (i.e., the CD-component), with some possessing dual heads, Table 12 presents the top 13 most frequent AB-components in terms of their type counts and Table 13 presents the most

---

17. It bears repeating that the analyzed list of 4KCW represents only 7.75% of all 4KCWs within the CWLs, while the analyzed list of 3KCWs represents 13.5% of all 3KCWs. Thus, it is highly probably that many more 4KCWs exist that are combinations of numbers and numerical units, but which are of lower frequencies (lemma frequencies > 10). It also bears noting that compound words that consist of three number kanji and a numerical unit/classifier (e.g., 三十一 + 日) are likely to far less frequent in occurrence compared to both a single number kanji and classifier (i.e., 2KCWs such as 一回 /ik-kai/ *one-time* [one + time]) and two number kanji and classifier (i.e., 3KCWs such as 三十分 30 minutes).

frequent 4KCWs, in terms of their token counts, for each of the most frequent AB-components.

TABLE 12. Top 13 most frequent AB-components of [AB]+[CD] 4KCWs according to type counts

AB	Gloss	Translation and explanation	Type count
当該	tō-gai	appropriate; relevant [appropriate + above-stated]	112
経済	kei-zai	economic; finance [pass thru; expire + settle (debt, etc.)]	88
自己	ji-ko	self; oneself [oneself + self]	82
生活	sei-katsu	living; life [life + lively]	79
国際	koku-sai	international [country + occasion; side]	78
社会	sha-kai	society, community [association + meeting; association]	76
一般	ip-pan	general, typical [one + general]	67
経営	kei-ei	business, management [pass thru; expire + occupation]	66
基本	ki-hon	fundamental, basic [foundation + base]	61
教育	kyō-iku	education; instruction [teach + raise]	58
政治	sei-ji	politics; government [politics + reign; rule]	58
生産	sei-san	production; manufacture [life; birth + product; yield]	58
地域	chi-iki	region [earth + region]	58

Highly consistent with the large-scale BCCWJ corpus from which the 4KCW database list has been derived, the most frequent AB-components are related to general areas of human activity, such as 経済 /kei-zai/ *economics*, 社会 /sha-kai/ *society*, 経営 /kei-ei/ *business* and 政治 /sei-ji/ *politics*. The most productive AB-component is the adjective 当該, which appears as the AB-component of 112 4KCWs within the database list, such as in 当該各号 /tō-gai-kaku-gō/ *relevant items* [[relevant + above-stated] + [each + item]] and 当該年度 /tō-gai-nen-do/ *relevant year(s)* [[relevant + above-stated] + [year + time]]. Apart from 当該, the other 12 most frequent AB-compounds are either nouns or VNs. For example, the second most frequent AB-component is the noun 経済, which appears as the AB-component of 88 4KCWs, such as 経済成長 /kei-zai-sei-chō/ *economic growth* [[expire + settle] + [become + long]] and 経済発展 /kei-zai-hat-ten/ *economic development* [[expire + settle] + [start from + unfold]]. The fourth ranked AB-component of 生活 is a VN, which appears as the AB-component of 79 4KCWs, such as 生活環境 /sei-katsu-kan-kyō/ *living environment* [[life + lively] + [ring + boundary]] and 生活習慣 /sei-katsu-shū-kan/ *lifestyle; living habits* [[life + lively] + [learn + accustomed to]].

TABLE 13. Most frequent [AB]+[CD] 4KCWs by token counts, for each of the most frequent AB-components

[AB]+[CD]	Gloss	Translation and explanation	Type count
当該各号	tō-gai-kaku-gō	relevant items [[relevant + above-stated] + [each + item]]	214
経済成長	kei-zai-sei-chō	economic growth [[expire + settle] + [become + long]]	689
自己責任	ji-ko-seki-nin	self-responsibility [[oneself + self] + [condemn + duty]]	356
生活環境	sei-katsu-kan-kyō	living environment [[life + lively] + [ring + boundary]]	822
国際社会	koku-sai-sha-kai	international society [[country + side] + [company + meet]]	786
社会主義	sha-kai-shu-gi	socialism [[association + meeting] + [main + meaning]]	563
一般会計	ip-pan-kai-kei	general accounting [[one + general] + [meeting + measure]]	473
経営戦略	kei-ei-sen-ryaku	management strategy [[expire + work] + [battle + outline]]	199
基本方針	ki-hon-hō-shin	basic policy [[foundation + base] + [direction + needle]]	839
教育訓練	kyō-iku-kun-ren	education + training [[teach + raise] + [instruct + practice]]	380
政治活動	sei-ji-katsu-dō	political activity [[politics + rule] + [lively + move]]	198
生産活動	sei-san-katsu-dō	production activity [[life + product] + [lively + move]]	281
地域社会	chi-iki-sha-kai	regional community [[earth + region] + [company + meet]]	1,007

Kobayashi, Yamashita, and Kageyama (2016, pp. 116–117) comment that nearly all [AB]+[CD] 4KCWs function as either nouns, VNs or ANs, as some of their examples in (18) illustrate.

(18)　[AB]+[CD] nouns
　　　財務大臣　/zai-mu-dai-jin/ *Finance Minister*
　　　　　　　　[[money + duties] + [big + retainer]]
　　　土地家屋　/to-chi-ka-oku/ *land and buildings*
　　　　　　　　[[soil + earth] + [house + roof]]
　　　[AB]+[CD] VN
　　　大学改革　/dai-gaku-kai-kaku/ *university reform*
　　　　　　　　[[big + learn] + [modify + reform]]
　　　意気消沈　/i-ki-shō-chin/ *depressed in spirits*
　　　　　　　　[[mind + spirit] + [extinguish + sink]]

[AB]+[CD] AN

利用可能　　/ri-yō-ka-nō/ *usable* [[benefit + use] + [can + ability]]
単純明快　　/tan-jun-mei-kai/ *simple and clear*
　　　　　　[[simple + pure] + [bright + pleasant]]

However, of the seven 4KCWs with [AB]+[CD] structures among the 20 most frequent by token frequency counts (Table 11), all are nouns apart from the one AN of 一生懸命 /is-shō-ken-mei/ *with utmost effort* [[one + life] + [depend + fate]]. Moreover, of the 4KCWs for each of the most frequent AB-components (Table 13), most are nouns, with just the three VNs of 教育訓練, 政治活動, and 生産活動.

Turning next to the CD-components of [AB]+[CD] 4KCWs, Table 14 presents the top ten most frequent CD-components in terms of their type counts and Table 15 presents the most frequent 4KCWs, in terms of their token counts, for each of the most frequent CD-components. Also highly consistent with the nature of corpora lexicons, the most frequent CD-components are also closely related to human activities. However, in contrast to the domain connotations of the AB-components, the most frequency CD-components by type counts primarily pertain to the notions of 関係 /kan-kei/ *relations*, 活動 /katsu-dō/ *activities*, 時間 /ji-kan/ *time* and 期間 /ki-kan/ *periods*, and 方法 /hō-hō/ *methods*, as well as 問題 /mon-dai/ *problems* and their 状況 /jō-kyō/ *situations* and 状態 /jō-tai/ *states*.

TABLE 14. Top 10 most frequent CD-components of [AB]+[CD] 4KCWs according to type counts

CD	Gloss	Translation and explanation	Type count
関係	kan-kei	relation; connection [connection + connection]	164
活動	katsu-dō	activity; action [lively + move]	156
以上	i-jō	… and upwards; beyond … [by means of + up]	154
時間	ji-kan	time; period [time + interval]	143
方法	hō-hō	method; process [way + method]	133
期間	ki-kan	period; term [period + interval]	124
主義	shu-gi	doctrine; -ism [main + meaning]	118
問題	mon-dai	problem; issue [ask + topic]	118
状況	jō-kyō	situation; circumstances [state + situation]	112
状態	jō-tai	state; condition [state + condition]	103

The most productive CD-component is the noun 関係, which appears as the CD-component of 164 [AB]+[CD] 4KCWs within the database list, such as in 人間関係 /nin-gen-kan-kei/ *human relations* [[human + space] + [connect + connect]] and 信頼関係 /shin-rai-kan-kei/ *relationship of mutual trust* [[faith + trust] + [connect + connect]]. The second most frequent

TABLE 15. Most frequent [AB]+[CD] 4KCWs by token counts, for each of the most frequent CD-components

[AB]+[CD]	Gloss	Translation and explanation	Type count
人間関係	nin-gen-kan-kei	human relations [[human + space] + [connect + connect]]	1,861
経済活動	kei-zai-katsu-dō	economic activity [[expire + settle] + [lively + move]]	519
必要以上	hitsu-yō-i-jō	more than necessary [[certain + need] + [by means of + up]]	504
労働時間	rō-dō-ji-kan	working hours [[labor + work] + [time + interval]]	790
応募方法	ō-bo-hō-hō	application method [[apply + recruit] + [way + method]]	292
一定期間	it-tei-ki-kan	fixed interval [[one + determine] + [period + interval]]	314
民主主義	min-shu-shu-gi	democracy [[people + main] + [main + meaning]]	1,102
環境問題	kan-kyō-mon-dai	environmental problem [[ring + boundary] + [ask + topic]]	900
実施状況	jis-shi-jō-kyō	implementation status [[real + perform] + [state + situation]]	326
健康状態	ken-kō-jō-tai	health condition [[healthy + ease] + [state + condition]]	326

CD-component is the VN of 活動, which is the only VN amongst the top ten CD-components. It is the CD-component of 156 4KCWs, such as 経済活動 /kei-zai-katsu-dō/ *economic activity* [[expire + settle] + [lively + move]] and 事業活動 /ji-gyō-katsu-dō/ *business activities* [[matter + business] + [lively + move]].

4.3. Secondary Morphological Structure of [ABC]+D 4KCWs

Reflecting the greater dominance of the primary [AB]+[CD] morphological structure for 4KCWs, the secondary structure of [ABC]+D only accounts for 12.1% of the 4KCW database list. Moreover, although this secondary structure closely parallels the primary [AB]+C morphological structure of 3KCWs, as noted earlier, where an additional morpheme is being attached to the end of an existing compound word, its coverage of only 12.1% stands in sharp contrast to the 77.1% prevalence of [AB]+C 3KCWs as the primary structure of 3KCWs. Moreover, further analyses of the D-additions of 4KCWs reveals that 26% are suffixes, which account for account for 61% of the [ABC]+D structures.

Table 16 presents the top ten most frequent D-additions to [ABC]+D 4KCWs by type counts and Table 17 presents the most frequent [ABC]+D 4KCWs, by token counts, for each of the most frequent D-additions.

TABLE 16. Top ten most frequent D-additions to [ABC]+D 4KCWs by type counts

D-addition	Meaning	Type count
等	etc.; and so forth	156
円	Japanese yen	152
人	*-er* person-indicating ending	147
条	article, clause, counter for articles	116
年	year	109
的	*-ic* AN ending	109
者	*-er* person-indicating ending	95
歳	age counter	76
間	between; interval	71
達	pluralizing ending	70

TABLE 17. Most frequent [ABC]+D 4KCWs by token counts, for each of the most frequent D-additions

[ABC]+D	Gloss	Translation and explanation	Type count
高齢者等	kō-rei-sha-tō	such as the elderly [[high + age + person] + pluralizer]	93
千五百円	sen-go-hyaku-en	1,500 yen [[thousand + five + hundred] + yen]	691
被相続人	hi-sō-zoku-nin	decedent [[cover + together + continue] + person]	297
第十二条	dai-jū-ni-jō	article 12 [[number + ten + two] + article]	636
二千一年	ni-sen-ichi-nen	2001 [[two + thousand + one] + year]	2,053
中長期的	chū-chō-ki-teki	mid-long term-ish [[middle + long + period] + -ic]	229
被保険者	hi-ho-ken-sha	insured person [[cover + protect + precipitous] + person]	1,013
二十四歳	ni-jū-yon-sai	24 years old [[two + ten + four] + years of age]	597
二十年間	ni-jū-nen-kan	20 year period [[two + ten + year] + interval]	381
主人公達	shu-jin-kō-tachi	protagonists [[main + person + public] + pluralizer]	15

In light of the clear parallels in terms of word-formation processes, it is most expedient to first compare the most frequent C-additions of [AB]+C 3KCWs (Table 6) with the most frequent D-additions of [ABC]+D 4KCWs (Table 16). While such comparisons reveal that five

morphemes are common to both lists (i.e., 的, 者, 等, 達, and 人), clearly, there are also differences in terms of their respective rankings. For instance, 的 is the most frequent C-addition, occurring in 873 [AB]+C 3KCWs, but it is only the sixth most frequent as a D-addition, occurring in 109 [ABC]+D 4KCWs, such as 中長期的 /chū-chō-ki-teki/ *mid-long term-ish* [[middle + long + period] + -ic]. However, in demonstrating that this morpheme attaches to both many 3KCWs and many 4KCWs, these results are highly consistent with Kobayashi et al.'s (2016, p. 127) observation, noted earlier, that 的 is a highly productive SJ affix that attaches to various bases. The most frequent D-addition is 等 which occurs in 156 4CKWs, such as 高齢者等 /kō-rei-sha-tō/ *such as the elderly* [[high + age + person] + pluralizer], while it is the third most frequent C-addition, occurring in 577 3KCWs. The largest shift in the respective frequency rankings for 3KCWs and 4KCWS is for 人, which is the third most frequent D-addition, occurring in 147 4KCWs, such as 被相続人 /hi-sō-zoku-nin/ *decedent* [[cover + together + continue] + person], as opposed to being the tenth most frequent C-addition, occurring in 227 3KCWs.

Comparing Tables 6 and 16 also reveals that five SJ morphemes are not common to both lists, but, highly congruent with earlier remarks about the likely frequency distributions of number kanji, these D-additions attach either solely or commonly to number kanji. Accordingly, it is not surprising to discover that the most frequent D-addition is 円 /en/ *Japanese yen currency*, which is a D-addition to 152 4KCWs, such as 千五百円 /sen-go-hyaku-en/ *1,500 yen* [[thousand + five + hundred] + yen] and of even larger sums, such as 五十万円 /go-jū-man-en/ *500,000 yen* [[five + ten + ten-thousand] + yen]. The fourth most frequent D-addition is 条 /jō/ *article, clause, counter for articles*, which occurs in 116 4KCWs, such as 第十二条 /dai-jū-ni-jō/ *article 12* [[number + ten + two] + article]. The fifth most frequent D-addition is 年 /nen/ *year*, which occurs in 109 4KCWs, such as 二千一年 /ni-sen-ichi-nen/ *2001* [[two + thousand + one] + year], while the eighth most frequent is 歳 /sai/ *age counter*, which occurs in 76 4KCWs, such as 二十四歳 /ni-jū-yon-sai/ *24 years old* [[two + ten + four] + years of age]. Although the ninth most frequent D-addition of 間 /kan/ *between; interval* also often combines with 3KCWs that involve numbers, such as 二十年間 /ni-jū-nen-kan/ *20 year period* [[two + ten + year] + interval], in such cases the C of the 3KCW invariably represents some time unit (such as minutes, days, months, and years). It can also attach to other kinds of 3KCWs, where the notion of between is spatial, such as 加盟国間 /ka-mei-koku-kan/ *between member states* [[add + alliance + country] + between].

4.4.   Tertiary Morphological Structure of A+[BCD] 4KCWs

As with the secondary structure of 4KCWs, the tertiary structure of A+[BCD] has also been considerably marginalized to just 1.9% of all 4KCWs structures, due to the marked prevalence of the 4KCW primary structure.   However, again, the parallels to the morphological structures of 3KCWs are present to the extent that the tertiary structure of A+[BCD] 4KCWs is similar to the secondary A+[BC] structure of 3KCWs, where an additional morpheme is being inserted at the beginning.   Moreover, the tendency seen with 3KCWs to derive longer compounds by appending a final morpheme as opposed to inserting an initial morpheme is also observed for the 4KCWs. As with the secondary structure of [ABC]+D 4KCWs, further analysis of the A-additions reveals that 32% are prefixes, which account for 75% of the A+[BCD] structures.

Table 18 presents the top ten most frequent A-additions to A+[BCD] 4KCWs by type counts and Table 19 presents the most frequent A+[BCD] 4KCWs, by token counts, for each of the most frequent A-additions.

TABLE 18. Top ten most frequent A-additions to A+[BCD] 4KCWs by type counts

A-addition	Meaning	Type count
約	approximately	84
各	each; every	46
総	gross, whole, general	24
同	same	22
新	new	16
全	all, whole	16
非	negation prefix	16
大	large; big	14
翌	the following; next	12
副	vice-; assistant	11

Also reflecting the close parallels in terms of word formation, there is again merit in comparing the most frequent A-additions for 3KCWs (Table 8) with the most frequent A-additions of 4KCWs (Table 18). Five morphemes are common to both lists (i.e., 大, 各, 新, 同, and 全), but the shifts in their respective ranking orders are generally not as pronounced as the shifts between the C-additions and D-additions to 3KCWs and 4KCWs, respectively.   However, in sharp contrast to 御 /o/ and /go/ *honorific prefix* being the most frequent A-addition for 3KCWs in terms of type counts, in the case of A+[BCD] 4KCWs, the most frequent A-addition is 約 /yaku/ *approximately,* which is an A-addition to 84 4KCWs, even though it is not amongst the top ten as an A-addition to 3KCWs.

TABLE 19. Most frequent A+[BCD] 4KCWs by token counts, for each of the most frequent A-additions

A+[BCD]	Gloss	Translation and explanation	Type count
約三十分	yaku-san-jip-pun	about 30 minutes [about + three + ten + minutes]	123
各市町村	kaku-shi-chō-son	each municipality [each + city + town + village]	113
総司令部	sō-shi-rei-bu	headquarters [general + official + orders + section]	134
同委員会	dō-i-in-kai	same committee [same + committee + member + meet]	116
新事業者	shin-ji-gyō-sha	new business person [new + thing + business + person]	94
全十二回	zen-jū-ni-kai	twelve times in total [all + ten + two + times]	37
非製造業	hi-sei-zō-gyō	nonmanufacturing sector [un + make + create + business]	115
大真面目	ō-majime	deadly serious [big + true + face + eye]	187
翌営業日	yoku-ei-gyō-bi	next working day [next + conduct + business + day]	23
副大統領	fuku-dai-tō-ryō	vice-president [vice + big + govern + territory]	67

As in both 約三十分 /yaku-san-jip-pun/ *about 30 minutes* [about + [three + ten + minutes]] and 約二百人 /yaku-ni-hyaku-nin/ *about 200 people* [about + [two + hundred + people]], 約 is typically inserted at the beginning of 3KCWs with [AB]+C structures, where the AB morphemes are numbers and the C-component is a numerical unit or classifier, such as minutes, people, and Japanese yen.

The second most frequent A-addition for 4KCWs is 各 /kaku/ *each; every*, which occurs in 46 4KCWs, such as in 各市町村 /kaku-shi-chō-son/ *each municipality* [each + [city + town + village]] and 各自治体 /kaku-ji-chi-tai/ *each municipality* [each + [[self + rule + body]]. Its ranking as the second most frequent A-addition is comparable to its ranking as the third most frequent A-addition for 3KCWs, which underscores the general productivity of this SJ morpheme as a prefix of both 3KCW and 4KCWs. Although not appearing within the top ten A-additions for 3KCWs, the third most frequent for 4KCWs is 総 /sō/ *gross, whole, general*, which occurs in 24 4KCWs, such as 総司令部 /sō-shi-rei-bu/ *headquarters* [general + [official + orders + section]] and 総事業費 /sō-ji-gyō-hi/ *total operating expenses* [gross + [matter + business + expenses]].

Of the four A-additions that function as negative prefixes (Kobayashi, Yamashita, and Kageyama, 2016), as noted earlier, only 非 /hi/ *negation* features within the top ten most frequent A-additions for 4KCWs,

even though it was not amongst the top ten for 3KCWs. It occurs in 16 4KCWs, such as 非製造業 /hi-sei-zō-gyō/ *nonmanufacturing sector* [un + [make + create + business]] and 非喫煙者 /hi-kitsu-en-sha/ *non-smoker* [non + [consume + smoke + person]].

4.5.   Other 4KCW Morphological Structures

Our analysis of the morphological structures of the 4KCW database reveals that a large majority (85.3%) have [AB]+[CD] structures, being the combination of two 2KCWs. The secondary and tertiary structures of 4KCWs involve one morpheme being added to an existing 3KCW, which together account for 14.0% of 4KCws However, as Table 10 also indicates, eight other morphological structures underlie a small percentage of 4KCWs. Accordingly, this section turns to present examples of those 4KCW structures.

For the 4KCWs, the first of these more marginal morphological structures is non-divisible (0.1%), which, as with the 3KCWs, is necessary to handle a small set of exceptions. The examples provided in (19) also illustrates that although the compound word's etymology and morphological structure are not clear, the meanings of the component morphemes are often related to the overall meaning.

(19)   Non-divisible
        炭水化物   /tan-sui-ka-butsu/ *carbohydrate*
                [coal + water + change + matter]
        不可思議   /fu-ka-shi-gi/ *mystery; unfathomable*
                [negative + can + think + debate]

The second of the marginal morphological structures is [A(CD)*]+[BCD] (0.1%), where the CD-component of an [ACD] 2KCW is omitted and the resultant A is attached to a related [BCD] 3KCW. This is also a form of clipping, as noted earlier, and, once again, this structure may appear to resemble superficially the A+[BCD] structure outlined above, to the extent that an A-component is being inserted before a [BCD] 3KCW. However, as with the [A(C*)]+[BC] structure of 3KCWs, the [A(CD)*]+[BCD] structure crucially hinges on the semantic relationship between the [ACD] and [BCD] 3KCWs, due to their shared CD-components, as the examples in (20) illustrates.

(20)   [A(CD*)]+[BCD] (with (*CD) omitted)
        小中学生   /shō-chū-gaku-sei/ *elementary and junior-high school students*
                [小 of 小学生 elementary school student + [中学生 junior-high school student]]
        土日曜日   /do-nichi-yō-bi/ *Saturday and Sunday*
                [土 of 土曜日 Saturday + [日曜日 Sunday]]s

The third of the marginal morphological structures is [A(D*)]+ [B(D*)]+[CD] (0.1%), where (D*) is omitted from both an [A(D*)] and a [B(D*)] 2KCW and the resultant A and B morphemes are inserted at the beginning of a CD 2KCW. As yet another example of a compound word formation that involves clipping, this structure also attests to the fact that the clipping process is a commonplace phenomenon. It is also essential to carefully differentiate this [A(D*)]+[B(D*)]+[CD] structure from the primary 4KCW structure of [AB]+[CD]. Although it may again potentially appear as if an [AB] 2KCW is being combined with a [CD] 2KCW, crucially, the A and B morphemes that are being inserted before the [CD] 2KCW here do not occur together as a 2KCW. This morphological structure also depends on the semantic connections between three 2KCWs due to the D component that is shared by all, as the examples in (21) highlight.

(21)   [A(D*)]+[B(D*)]+[CD] (with both (*D) omitted)
        陸海空軍   /riku-kai-kū-gun/ *land, sea and air forces*
                   [land + sea + air + troops] [陸 of 陸軍 land forces + 海 of 海軍
                   navy + [空軍 air force]]
        農林漁業   /nō-rin-gyo-gyō/ *agriculture, forestry and fishing*
                   [farm + forest + fish + industry] [農 of 農業 agriculture + 林 of
                   林業 forestry + [漁業 fishing industry]]

The fourth of the marginal morphological structures is A+B+C+D (0.1%), as the concatenation of four morphemes that together constitute a set of things, with the examples in (22) being prototypical.

(22)   A+B+C+D
        春夏秋冬   /shun-ka-shū-tō/ *four seasons*
                   [spring + summer + autumn + winter]
        喜怒哀楽   /ki-do-ai-raku/ *human emotions*
                   [joy + anger + grief + pleasure]

The fifth of the marginal morphological structures is phonological transcription (当て字) (0.1%). As explained earlier for the 3KCWs structures, there are also 4KCWs where the kanji are being used conventionally to represent the word's syllables, as in (23).

(23)   Phonological transcriptions (当て字)
        滅茶滅茶   /me-cha-me-cha/ *disorderly, absurd; excessive*
                   [destroy + tea + destroy + tea]
        無理矢理   /mu-ri-ya-ri/ *forcibly; against one's will*
                   [nothing + reason + arrow + reason]

The sixth of the marginal morphological structures for 4KCWs is [AB]+C+D (0.0%), where C and D morphemes are being attached to an [AB] 2KCW. This structure should also be distinguished from the primary morphological structure of [AB]+[CD], because, as with the A and

B morphemes inserted before a CD 2KCW in the [A(D*)]+[B(D*)]+[CD] structure, the C and D morphemes that are attached to form [AB]+C+D structures do not occur together as an independent 2KCW. The AB components of the 4KCWs that conform to this structure are number kanji and the C morpheme is 箇 /ka/ *counter for articles*, as the examples in (24).

(24)　　[AB]+C+D
　　　　十二箇月　　/jū-ni-ka-getsu/ *12 month (period)*
　　　　　　　　　　[[ten + two] + counter + month]
　　　　十一箇国　　/jū-ichi-ka-koku/ *11 countries*
　　　　　　　　　　[[ten + one] + counter + country]

　　The seventh of the marginal morphological structures is monomorphemic words (熟字訓) (0.0%). Again, in contrast to phonological transcriptions, although there is usually no phonological correspondence between the elements of the graphematic representation and the compound word pronunciation, the meanings of the component kanji usually relate to the word's overall meaning, which is the case with the example in (25).

(25)　　Monomorphemic words (熟字訓)
　　　　再従兄弟　　/haitoko/ *second cousin*
　　　　　　　　　　[again + accompany + elder brother + younger brother]

　　The eighth and final of the marginal morphological structures for 4KCWs is [A(D*)]+[BCD] (0.0%), where the D-component of an AD 2KCW is omitted and inserted at the beginning of an BCD 3KCW. It is also important to distinguish this structure from both the secondary structure of A+[BCD] and from the second of the more marginal structures of [A(CD)*]+[BCD] with the (CD) element of the A(CD) 3KCW omitted. As with the second of the marginal structures, the distinction is well motivated based on the semantic connection between the D-component of the [A(D)*] 2KCW and the D component of the [BCD] 3KCW, as the example in (26) illustrate.

(26)　　[A(D*)]+[BCD] (with (*D) omitted)
　　　　産婦人科　　/san-fu-jin-ka/ *maternity and gynaecology*
　　　　　　　　　　[産 of 産科 obstetrics + [婦人科 gynaecology]]

5.　Concluding Remarks

This paper has outlined the construction of two new databases of 3KCWs and 4KCWs, as key components for a larger database project concerned with Japanese lexical properties (Joyce, Hodošček, and Masuda, 2017; Joyce, Masuda, and Ogawa, 2014). More specifically, this paper has focused on describing the results of analysing the extracted data-

base lists according to the morphological structures that underlie the 3KCWs and 4KCWs, respectively. The results provide tangible quantitative indications of the degrees to which the dominant morphological structures differ for 3KCWs and 4KCWs (Kageyama and Saito, 2016; Kobayashi, Yamashita, and Kageyama, 2016; Shibatani, 1990; Tamamura, 1984; 1985).

In the case of the 3KCW database, although eight structures were identified in total, the analysis results clearly show that just two morphological structures underlie the vast majority (98.4%) of the 23,046 3KCWs. Moreover, although both structures involve adding a morpheme to an existing 2KCW, the results also reveal a striking preference for attaching an morpheme to the end of an existing 2KCW, such that the primary morphological structure of [AB]+C accounts for 77.1% of the 3KCWs. In comparison, the secondary structure of A+[BC], where the additional morpheme is added to the beginning of a 2KCW, only accounts for 21.3% of the 3KCWs. In sharp contrast to the results for the 3KCWs, in the case of the 4KCW database, although 11 structures were identified in total, the analysis results indicate that the dominant morphological structure of 4KCWs is overwhelmingly [AB]+[CD], where two 2KCWs are combined by compounding processes, and which account for 85.3% of the 23,159 4KCWs. Notwithstanding the pervasive nature of the primary structure, still, a considerable number of 4KCWs are formed by adding a morpheme to an existing 3KCW, where a marked preference for attachment to the end of compound words is also observed. Thus, at much reduced proportions compared to 3KCWs, for 4KCWs, the secondary structure is [ABC]+D, which accounts for 12.1%, and the tertiary structure is A+[BCD], which accounts for 1.9%.

Taken together, the results of analyzing the morphological structures of both database lists unquestionably underscore the immense significance of 2KCWs within the Japanese lexicon, not only as words in their own right, but as the component elements of longer compound words (Joyce, 2011; Joyce, Hodošček, and Masuda, 2017; Joyce, Masuda, and Ogawa, 2014; Kobayashi, Yamashita, and Kageyama, 2016; Nomura, 1975; 1988). Overall, these findings are entirely consistent with the morphographic nature of kanji (Joyce, 2011; Kobayashi, Yamashita, and Kageyama, 2016) because they vividly highlight how the concatenation of kanji in graphematically representing the vast majority of Japanese compound words is primarily the province of the morphological processes that underlie the formation of Japanese compound words. That is, while there are undeniably a limited number of exceptions, such as the non-divisible, phonological transcription and monomorphemic structures, the surface graphematic forms of most Japanese compound words conform to the morphographic principle (Joyce, 2011). However, kanji are associated with both NJ and SJ morphemes, many with multiple NJ and SJ allomorphs, and the status of those morphemes—as either free, bound or affixes—is often context-dependent. Accordingly, the present analyses

of the rich morphological structures of Japanese compound words can potentially further elucidate the intricate nature of the morphographic principle in the case of the JWS; a topic that undoubtedly warrants greater attention from the perspective of writing systems research.

As indicated earlier, the task of analyzing the morphological structures of 3KCWs and 4KCWs has been greatly facilitated by the fact that, in fundamentally conforming to the morphographic principle, their structures are generally highly transparent. However, as also acknowledged, reflecting the productive nature of SJ morphemes, some compound words are conceivably open to alternative interpretations, such as 農業者 *agricultural worker*. Accordingly, even though the most plausible interpretation of 農業者 is to regard it as an example of the [AB]+C structure, we are also planning to conduct studies to obtain native-speaker rankings related to the psychological validity and credibility of the morphological structures. Such studies will also investigate the extent to which semantic shifts in the meanings of compound words might influence native-speaker interpretations of their morphological structures. For instance, although the meanings of the constituent morphemes are clear and the morphological structure of 新幹線 /shin-kan-sen/ [new + [trunk + line]] is unquestionably A+[BC], the compound word's contemporary meaning of *bullet train* represents a substantial semantic shift.

Moreover, the conducted analyses of the morphological structures of both 3KCWs and 4KCWs are essential for preparing to conduct various visual word recognition experiments to further investigate Kobayashi et al.'s (2016, p. 129) claims that kanji facilitate meaning comprehension, which have significant implications for the organization of morphological information within the mental lexicon. Joyce (2002; 2004) and Masuda and Joyce (2018) have already conducted a series of psycholinguistic experiments that have employed the constituent-priming paradigm to examine lexical-decision task responses to 2KCWs. As those studies have generally observed robust patterns of facilitated reaction times due to the prior presentation of the constituent kanji, across a variety of conditions, including very brief stimulus onset asynchrony intervals, the natural next steps are to extend this experimental approach to investigate the recognition processes of 3KCWs and 4KCWs. To that aim, the analyses of their morphological structures will be invaluable in terms of designing various experimental conditions and selecting suitable stimuli.

As already noted, these 3KCW and 4KCW databases have been compiled as new components of a larger database project concerned with various Japanese lexical properties (Joyce, Hodošček, and Masuda, 2017; Joyce, Masuda, and Ogawa, 2014; Masuda and Joyce, 2005; Masuda, Joyce, et al., 2014). In being extracted from Joyce, Hodošček, and Nishina (2012) CWLs, which were, in turn, extracted from the BCCJW, both database lists have automatically inherited a number of valuable data-fields, such as word class, lexical strata, token frequencies of lemma

and orthographic base forms and pronunciation(s). In addition to those inherited data-fields, naturally, the work of analysing the morphological structures has itself generated a number of additional fields beyond just assigning a structure category, such as identifying all possible graphematic overlaps and counts related to morphological family sizes. These will all be checked as the work of integrating these new database components within the larger database progresses and as further database components are developed in the future, such as the planned analyses of the morphological structures of five-kanji compound words.[18] Thus, the present analyses of the morphological structures of both 3KCWs and 4KCWs represent a significance contribution to the larger database project of mapping out various Japanese lexical properties.

References

Agency for Cultural Affairs [文化庁] (2010). "常用漢字表 [Jōyō kanji list]." In: URL: http://kokugo.bunka.go.jp/kokugo_nihongo/joho/kijun/naikaku/pdf/joyokanjihyo_20101130.pdf.

Backhouse, A.E. (1984). "Aspects of the graphological structure of Japanese." In: *Visible Language* 18.3, pp. 219–228.

Igarashi, Yuko (2007). "The changing role of katakana in the Japanese writing system: Processing and pedagogical dimensions for native speakers and foreign learners." PhD thesis. University of Victoria, British Columbia, Canada.

Joyce, Terry (2002). "Constituent-morpheme priming: Implications from the morphology of two-kanji compound words." In: *Japanese Psychological Research* 44, pp. 79–90.

———— (2004). "Modeling the Japanese mental lexicon: Morphological, orthographic and phonological considerations." In: *Advances in Psychological Research: Volume*. Ed. by S.P. Shohov. Vol. 31. Hauppauge, NY: Nova Science, pp. 27–61.

———— (2011). "The significance of the morphographic principle for the classification of writing-systems." In: *Written Language & Literacy* 14.1, pp. 58–81.

Joyce, Terry, Bor Hodošček, and Hisashi Masuda (2017). "Constructing an ontology and database of Japanese lexical properties: Handling the orthographic complexity." In: *Written Language & Literacy* 20.1, pp. 27–51.

18. Although longer Japanese compound words are frequently attested, as Kobayashi, Yamashita, and Kageyama (2016, pp. 114–115) observe, as longer compound words invariably involve recursive combinations of existing shorter compound words. Thus, the returns from analyzing beyond five-kanji compound words are likely to be rather limited.

Joyce, Terry, Bor Hodošček, and Kikuko Nishina (2012). "Orthographic representation and variation within the Japanese writing system: Some corpus-based observations." In: *Written Language & Literacy* 15.2, pp. 254–278.

Joyce, Terry and Hisashi Masuda (2018). "Introduction to the multi-script Japanese writing system and word processing." In: *Writing systems, reading processes, and cross-linguistic influences: Reflections from the Chinese, Japanese and Korean languages*. Ed. by Hye Pae. Vol. 7. Bilingual Processing and Acquisition. Amsterdam: John Benjamins, pp. 179–199.

———— (2019). "On the notions of graphematic representation and orthography from the perspective of the Japanese writing system." In: *Written Language & Literacy* 22.2, pp. 248–280.

Joyce, Terry, Hisashi Masuda, and Taeko Ogawa (2014). "Jōyō kanji as core building blocks of the Japanese writing system: Some observations from database construction." In: *Written Language & Literacy* 17.2, pp. 173–194.

Kageyama, Taro and Michiaki Saito (2016). "Vocabulary strata and word formation processes." In: *Handbook of Japanese lexicon and word formation*. Ed. by Taro Kageyama and Hideki Kishimoto. Vol. 3. Handbooks of Japanese Language and Linguistics. Boston, Berlin: Walter de Gruyter, pp. 11–50.

Kess, Joseph F. and Tadao Miyamoto (1999). *The Japanese mental lexicon: Psycholinguistics studies of kana and kanji processing*. Amsterdam: John Benjamins.

Kobayashi, Hideki, Kiyo Yamashita, and Taro Kageyama (2016). "Sino-Japanese words." In: *Handbook of Japanese lexicon and word formation*. Ed. by Taro Kageyama and Hideki Kishimoto. Vol. 3. Handbooks of Japanese Language and Linguistics. Boston, Berlin: Walter de Gruyter, pp. 93–131.

Konno, Shinji [今野真二] (2013). 正書法のない日本語 *[The Japanese language lacks orthography]*. 東京 [Tokyo]: 岩波書店 [Iwanami Shoten].

Lurie, David B. (2012). "The development of writing in Japan." In: *The shape of script: How and why writing systems change*. Ed. by S.D. Houston. Santa Fe, NM: School for Advanced Research Press, pp. 159–185.

Maekawa, Kikuo et al. (2013). "Balanced corpus of contemporary written Japanese." In: *Language Resources and Evaluation*, pp. 1–27.

Masuda, Hisashi and Terry Joyce (2005). "A database of two-kanji compound words featuring morphological family, morphological structure, and semantic category data." In: *Corpus Studies on Japanese Kanji*. Ed. by Katsuo Tamaoka. Vol. 10. Glottometrics. Tokyo, Japan: Hituzi Syobo, Lüdenschied, Germany: RAM-Verlag, pp. 30–44.

———— (2018). "Constituent-priming investigations of the morphological activation of Japanese compound words." In: *Writing systems, reading processes, and cross-linguistic influences: Reflections from the Chinese,*

Japanese and Korean languages. Ed. by Hye Pae. Amsterdam: John Benjamins, pp. 221–244.

Masuda, Hisashi and Terry Joyce (2019). "A database of three-kanji compound words in Japanese, with particular focus on their morphological structures." Poster presentation given as the *'Diversity of writing systems: Embracing multiple perspectives', 12th International Workshop on Written Language and Literacy*, Faculty of Classics, Cambridge University, UK.

Masuda, Hisashi, Terry Joyce, et al. (2014). "A database of semantic transparency ratings for two-kanji Japanese compound words." Poster presentation given at *'Orthographic Databases and Lexicons': 9th International Workshop on Writing Systems and Literacy*, University of Sussex, Brighton, UK.

Miller, Laura (2011). "Subversive script and novel graphs in Japanese girls' culture." In: *Language & Communication* 31.1, pp. 16–26.

Nomura, Masaaki [野村雅昭] (1975). "四字漢語の構造 [The structure of four-kanji Sino-Japanese words]." In: 電子計算機による国語研究 *[Studies in Computational Linguistics]* 7, pp. 36–80.

——— (1988). "二字漢語の構造 [The structure of two-kanji Sino-Japanese words]." In: 日本語学 *[Japanese Studies]* 7.5, pp. 44–55.

Robertson, Wesley C. (2015). "Orthography, foreigners, and fluency: Indexicality and script selection in Japanese manga." In: *Japanese Studies* 35.2, pp. 205–222.

——— (2017). "He's more katakana than kanji: Indexing identity and self-presentation through script selection in Japanese manga (comics)." In: *Journal of Sociolinguistics* 21.4, pp. 497–520.

Shibatani, Masayoshi (1990). *The Languages of Japan*. Cambridge, UK: Cambridge University Press.

Shinmura, Izuru [新村出], ed. (1995). 広辞苑 *[Japanese dictionary]*. 5th ed. 東京 [Tokyo]: 岩波書店 [Iwanami Shoten].

——— ed. (2008). 広辞苑 *[Japanese dictionary]*. 6th ed. 東京 [Tokyo]: 岩波書店 [Iwanami Shoten].

Smith, Janet S. (Shibamoto) (1996). "Japanese writing." In: *The world's writing systems*. Ed. by Peter T. Daniels and William Bright. New York: Oxford University Press, pp. 209–217.

Smith, Janet S. (Shibamoto) and David L. Schmidt (1996). "Variability in written Japanese: Towards a sociolinguistics of script choice." In: *Visible Language* 30.1, pp. 47–71.

Tamamura, Fumio [玉村文郎] (1984). 日本語教育指導参考書12: 語彙の研究と教育（上）*[Japanese language education reference guides 12: Lexical research and education 1]]*. 東京 [Tokyo]: 大蔵省印刷局 [Ookurashou Insatsukyoku].

——— (1985). 日本語教育指導参考書13: 語彙の研究と教育（下）*[Japanese language education reference guides 13: Lexical research and education 2]]*. 東京 [Tokyo]: 大蔵省印刷局 [Ookurashou Insatsukyoku].

Taylor, Insup and M. Martin Taylor (2014). *Writing and literacy in Chinese, Korean and Japanese*. Vol. 14. Studies in Written Language and Literacy. Amsterdam: John Benjamins.

Tranter, Nicolas (2008). "Nonconventional script choice in Japan." In: *International Journal of the Sociology of Language* 192, pp. 133–151.

A Modular Theoretic Approach to the Japanese Writing System: Possibilities and Challenges

Keisuke Honda

Abstract. The Modular Theory of Writing Systems (MT) provides a three-module model of the correspondence between the elements of a script and the properties of a language at the level of individual words. Characterised by its non-derivational linguistic approach, MT has the potential to develop into a general theory of script-to-language relationship in any type of writing system. However, it is currently focused on the analysis of modern alphabetic systems, with little regard for non-alphabetic systems. To examine the theory's compatibility with a typologically wider range of writing systems, the present paper discusses the functional aspects of the present-day Japanese writing system within the MT framework. This system offers a good testing ground because it makes a mixed use of logographic, moraic and alphabetic scripts. The discussion highlights the possibilities and challenges of current MT and presents some proposals to increase its applicability to non-alphabetic systems.

Introduction

Writing allows us to communicate linguistic messages through graphic representations in a conventional and systematic way (Gelb, 1963, pp. 11–20; DeFrancis, 1989, pp. 4–6; Coulmas, 2003, pp. 1–17; Rogers, 2005, pp. 2–4; Sampson, 2015, pp. 18–39; Daniels, 2018, pp. 156–157). Each writing system enables this function by pairing a particular script with a specific language according to a unique set of conventions. Despite their rich diversity, the world's writing systems show important similarities—as well as differences—in the way they relate the elements of a script to the properties of a language (e.g., Justeson, 1976, pp. 58–

Keisuke Honda   ⓘD 0000-0003-4228-5406
Imperial College London, Centre for Languages, Culture and Communication
South Kensington Campus, London SW7 2AZ, United Kingdom

University of Oxford, Oxford University Language Centre
12 Woodstock Road, Oxford OX2 6HT, United Kingdom
E-mail: kdhonda@gmail.com

76). An important task of grapholinguistics, then, is to develop a theoretical framework for describing and explaining this relationship in and across writing systems.[1]

It is in this context that the present paper focuses on the Modular Theory of Writing Systems (MT: Neef, 2012; 2015). MT provides a general model of script-to-language relationship at the level of individual words, conceptualised in terms of three modules called *language system*, *graphematics* and *systematic orthography* (Section 1). This model is built on two important assumptions that distinguish it from previous models of writing systems. The first assumption is that every writing system is constructed on an abstract language system (Neef, 2012, p. 4; 2015, p. 709). This notion opens the way for a uniform linguistic analysis of writing systems without viewing them as a surrogate of concrete speech (cf. Saussure, 1983/1916, p. 45; Bloomfield, 1933, p. 21) or as autonomous sign systems (cf. Vachek, 1973, pp. 14–17; Harris, 1995, pp. 56–63). The second assumption is that such a linguistically based analysis of writing systems should be based on declarative descriptions of the underlying language systems (Neef, 2015, p. 709). This paradigm accounts for properties of linguistic structure in terms of well-formedness conditions applicable to a single level of representation, instead of derivational rules converting one level of representation into another (cf. Chomsky and Halle, 1968, p. 49; Chomsky, 1970, pp. 287–294; Sproat, 2000, pp. 18–19). With this non-derivational linguistic approach, MT has the potential to expand into a general theory of how words are written in different writing systems (Section 1.3).

Crucially, however, the current MT model has a fairly limited scope of application. It draws almost entirely on observations of modern alphabetic systems like German and English, where characters and character combinations relate primarily to individual vowels and consonants.[2] In other words, MT makes virtually no mention of non-alphabetic systems, which may be either phonographic or highly lo-

------

1. Neef (2015, p. 711) defines grapholinguistics as "[t]he linguistic sub discipline dealing with the scientific study of all aspects of written language". Similarly, Haralambous (2020, p. 12), states that this field of research "aims to study aspects of language that are particular to its written representation, at all levels of linguistics". As is evident from the title of the present volume and its preceding conference, the term 'grapholinguistics' is becoming increasingly accepted as an alternative to other terms such as 'grammatology', 'graphology', 'graphemics', 'graphonomy' and 'writing systems research'. See Daniels (2018, pp. 4–5) for an overview of the various designations given to the study of writing, writing systems and written language.

2. The present paper uses 'alphabetic' to refer to any writing system based on both vocalic and consonantal segments. Some might prefer 'segmental', which appears to specify the type of underlying phonological unit. However, this alternative term is too broad because it covers all subtypes of segment-based systems without reference to the presence or absence of vocalic signs or the spatial arrangement of segmental signs (e.g., compare the writing systems of Finnish, Arabic, Hindi and Korean; for

gographic in nature (Section 1.2). This is a major drawback given the prevalence of such systems throughout the history of writing around the world. Hence, despite its designation as a "theory of writing systems" (Neef, 2015, p. 708), MT in its present form is effectively a theory of alphabetic systems. Therefore, it remains an open question whether it can actually be expanded into a full-fledged theory of script-to-language relationship across different types of writing systems.

This paper aims to address the above question through a partial analysis of the current Japanese writing system. As widely documented, this system employs a mixture of four main scripts that function as typologically distinct sets of written signs: logographic *kanji* (漢字), moraic *hiragana* (平仮名) and *katakana* (片仮名), and alphabetic *rōmaji* (ローマ字) (e.g., Smith, 1996, pp. 209–213; Sasahara, 2001, pp. 704–705; Honda, 2012, pp. 38–71; Taylor and Taylor, 2014, pp. 271–302). As such, it serves as a useful test case for examining the adaptability of MT to non-alphabetic systems. Through a discussion of the main characteristics of the Japanese writing system, the present paper seeks to highlight the possibilities and challenges of the current MT model. The discussion is organised as follows. Section 1 introduces the key concepts of MT. Section 2 discusses their applicability to the analysis of how the four scripts are used to write words in Japanese. Section 3 examines the notions of logography and logographic systems assumed in MT. Section 4 summarises the discussion and draws conclusions.

1.   Key Concepts of MT

As already mentioned, current MT is essentially a theory of alphabetic systems, where the elements of a script relate mainly to the segmental level of phonology. It is built on specific assumptions and claims about the formal and functional elements of writing systems (Section 1.1), the distinction between phonographic and logographic systems (Section 1.2), and the architecture of alphabetic systems (Section 1.3).

1.1.   Formal and Functional Elements

Every writing system employs a certain number of discrete graphic marks. Each mark can take a variety of similar but different shapes in written, printed, or electronically displayed texts. Thus, one can speak

---

discussions, see Faber, 1992, pp. 118–123, and Gnanadesikan, 2017, pp. 19–31). While a more accurate description would be obtained by adopting a combinatorial term like 'fully vowelled linear segmentary' (Gnanadesikan, 2017, p. 28), this option makes it difficult to refer to all non-alphabetic systems as a single class.

of a set of abstract forms embodied by a larger set of concrete shapes. In MT terminology (Neef, 2015, pp. 711–713), *script* and *character* respectively denote any such abstract set and each member thereof, whereas *font* and *glyph* denote their concrete counterparts.[3] This is illustrated by the lower case Roman script in (1); the pipes | | enclose each character of the script, and the arrow shows the character's correspondence to the glyphs of different fonts on the right side.

(1)   Script : Character :: Font : Glyph
      |a| →   a,   a,   a,   a,   *a*,   *α*,   …
      |b| →   b,   b,   b,   b,   *b*,   *ℓ*,   …
      |c| →   c,   c,   c,   c,   *c*,   *c*,   …
       ⋮
      |z| →   z,   z,   z,   z,   *z*,   *ʒ*,   …

Importantly, MT views scripts and characters (as well as fonts and glyphs) as purely formal elements of writing. For a script to function as a writing system, it must be paired with a particular language in a systematic way. At the basis of this pairing is a conventional association between characters or character combinations and different properties of the language in question. In the English writing system, for example, characters are associated with phonological units (e.g., |p| → [p]), morphosyntactic units (e.g., |\$| → DOLLAR) or syntactic information (e.g., |?| → 'interrogative'). Using a dyadic model of signs (Saussure, 1983/1916, pp. 99–100), one may speak of *written signs*, each comprising a character or character combination as the signifier and a linguistic property or a piece of linguistic information as the signified.[4] MT distinguishes four types of written signs based on their signifieds (Neef, 2015, p. 711). The present paper assumes this classification with a partly modified terminology as shown below, where the angle brackets <> enclose a written sign: *phonographs* correspond to phonological units (2a), *logographs* to morphosyntactic units (2b), *ciphers* to numbers (2c), and *punctuation marks* to information about linguistic structure (2d).[5]

----

3. As 'font' is conventionally associated with typography, 'hand' might be a better alternative for referring to any set of glyphs used in handwriting (Douglas, 2017, pp. 5–6).

4. Defined this way, the notion of 'written sign' is comparable with the semiotic reinterpretation of 'grapheme' proposed by Meletis (2019, p. 9–10). These and related concepts and terms require further discussion in the future.

5. Neef (2015, p. 712) uses 'letters' and 'logographs' to refer to (2a) and (2b), respectively. This paper adopts 'phonographs' for the first class instead, as 'letters' are conventionally restricted to the signs of phonological segments employed in alphabetic and consonantal systems (Sampson, 2015, pp. 10–11). Besides, for both (2a) and (2b), the *-graph* ending is preferred over the original *-gram* ending because only the former can be used in their derived forms (e.g., phonographic versus *phonogram-

(2)   Written signs
      a.   <p> : |p| → [p]
      b.   <$> : |$| → DOLLAR
      c.   <1> : |1| → ONE
      d.   <?> : |?| → 'interrogative'

Using these notions, MT distinguishes three aspects of writing systems. The first two belong to the formal aspect, one being concrete (i.e., glyphs and fonts) and the other being more abstract (i.e., characters and scripts). The third one is the functional aspect, where graphic signifiers are linked with linguistic signifieds (i.e., written signs). Each of these aspects is studied in different subfields of grapholinguistics: *typography* (3a) and *graphetics* (3b) are concerned with the formal aspects of writing systems, and MT with their functional aspects (3c) (Neef, 2015, p. 711).

(3)   Subfields of grapholinguistics
      a.   Typography   concerns glyphs and fonts
      b.   Graphetics   concerns characters and scripts
      c.   MT   concerns written signs

1.2.   Phonographic and Logographic Systems

MT adopts a traditional distinction of two broad types of writing systems according to the main type of written signs used therein (Neef, 2015, p. 713). In this scheme, a writing system is described as being *phonographic* if the characters principally function as phonographs, or *logographic* if they mainly operate as logographs (e.g., Sampson, 2015, pp. 24–26).[6]

Phonographic systems are further divided into subtypes according to the main type of phonological unit represented by the phonographs. While different studies use different classifications and terminologies, common labels include full segmental or alphabetic (e.g., English), consonantal or abjad (e.g., Arabic), alphasyllabic or abugida (e.g., Hindi),

---

mic). It may also be possible to treat ciphers (2c) as a subclass of logographs because they are associated with numerals as morphosyntactic units (e.g., ONE) rather than numerical concepts (e.g., 'lowest cardinal number'); however, this treatment requires further elaboration because some numerical notation systems operate in notably different ways from glottographic (i.e., language-based) notations (Pettersson, 1996, pp. 798–805; Sproat, 2000, p. 198). The present classification also needs to be refined to deal with other types of non-glottographic signs such as semantic classifiers and ideographs attested across writing systems.

    6. Some studies use the labels *cenemic* and *pleremic* to refer to these types of writing systems (Haas, 1983, p. 16). For a general overview of writing system typologies, see Daniels (1996, pp. 8–10) and Joyce and Borgwaldt (2011, pp. 1–6). For discussions, see Sproat (2000, pp. 132–144), Rogers (2005, pp. 269–279), Sampson (2015, pp. 18–39) and Joyce (2016, pp. 288–297).

moraic or core syllabic (e.g., Cherokee), full syllabic (e.g., Modern Yi), and so on.[7] Logographic systems, on the other hand, are considered to have only one subtype, namely morphemic systems employing mainly morphographs or the signs of individual morphemes (Hill, 1967, p. 93). This is summarised schematically in (4); the morphemic nature of logographic systems will be discussed in Section 3.

(4)   Phonographic and logographic systems
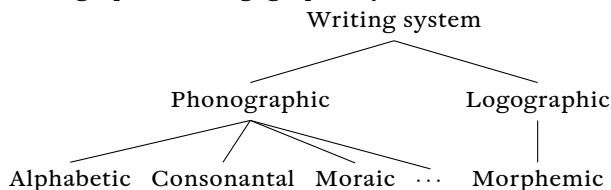
1.3.   Three Modules

As already mentioned, the current MT model is primarily concerned with the script-to-language relationship within words in modern alphabetic systems. This relationship is captured at a single level of abstraction, namely in terms of the correspondence between characters (as opposed to glyphs) or character combinations and phonological segments (as opposed to phones). Regarding the latter, MT's non-derivational linguistic approach entails an exclusive focus placed on the surface (as opposed to underlying) phonological representation. This is shown in (5), exemplified by the relationship between the written and phonological forms of the English word *writing*.

(5)   Script-to-language relationship for <writing>

Typography	writing		*Phonetics*
MT	|writing| →	[ˈɹaɪtɪŋ]	*Surface phonology*
		/raɪt+ɪŋ/	*Underlying phonology*
		WRITE+ing	*Morphology*
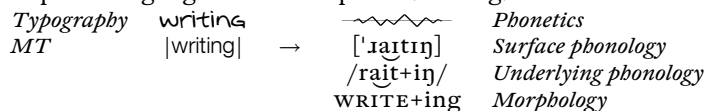
   MT assumes three modules to explain this relationship. The first module is the *language system*, which provides the foundation for each writing system to function as a language-based sign system (Neef, 2012, p. 4; Neef, 2015, pp. 709–711). Adopting the structural and generative conception of language, MT distinguishes two parts for each language system: *grammar*, which captures the regular aspects of a language, and *lexicon*, which covers all irregular properties in the same language. More specifically, grammar comprises phonology, morphology,

---

7. For a recent review and discussion of typologies of phonographic systems, see Sproat (2000, pp. 131–144), Ratcliffe (2001, pp. 3–6), Buckley (2018, pp. 32–46) and Gnanadesikan (2017, pp. 16–30).

semantics and syntax, whereas lexicon defines morphemes as the arbitrary associations of forms and meanings.[8]

A language system enables another module called *graphematics* to provide possible written representations of individual words in that language (Neef, 2012, pp. 4–6; Neef, 2015, pp. 713–714). This second module defines all conventional associations between characters or character combinations and phonological segments permitted in the writing system. If the graphematics allows more than one way to represent the same phonological segment, it generates a set of theoretically possible spellings for each word containing that segment. In MT terminology, this set is known as the *graphematic solution space* (Neef, 2012, pp. 10–11; Neef, 2015, p. 716). To illustrate, German [vɑl] denotes several distinct meanings including 'whale', 'choice' and different place names (Neef, 2012, pp. 10–11). It is possible to spell this phonological form in a number of different ways because the German graphematics permits multiple characters and character combinations to represent its constituent segments (6).

(6)   German graphematics for [v], [ɑ], [l][9]
     a. <w> ∨ <v> ∨ <vh> → [v]
     b. <a> ∨ <aa> ∨ <ah> → [ɑ]
     c. <l> ∨ <ll> ∨ <lh> → [l]

Consequently, there is a large graphematic solution space for each of the homophonous words. It contains both attested spellings and unattested but theoretically possible spellings (7).

(7)   German graphematic solution space for [vɑl]
     <wal> <waal> <wahl> <whal> <whaal> <whahl> <walh> …
     <val> <vaal> <vahl> <vhal> <vhaal> <vhahl> <valh> …

In actuality, however, different spellings are used by convention to distinguish between the homophones (8). In MT, this is explained by the third module called *systematic orthography*, which prescribes how to spell individual words correctly within the limitation of the graphematic solution space (Neef, 2012, pp. 11–13; 2015, pp. 715–718). These constraints are 'systematic' in the sense that they apply to particular layers

---

8. MT assumes that a writing system can refer to any part of the grammar and lexicon of a language. This point is exemplified by the spellings of French [ɛme] in different inflections (e.g., <aimer> 'love-INF' versus <aimez> 'love-3.PL'), which presuppose reference to both the phonological and morphosyntactic aspects of the forms in question (Neef, 2015, p. 710).

9. Throughout this paper, the disjunction symbol <∨> is used to indicate the presence of multiple items on either side of graphematic correspondence. There can be two or more characters or character combinations associated with a single linguistic property, or two or more linguistic properties associated with a single character or character combination.

of the vocabulary. For example, an analysis of German spelling justifies a constraint that <aa> cannot be used to represent [ɑ] (8a,b) except in foreign proper names (8c,d).

(8) German orthographic forms for [vɑl] homophones

    a. <Wal> 'whale'

    b. <Wahl> 'choice'

    c. <Waal> 'River Waal in the Netherlands'

    d. <Vaal> 'River Vaal in South Africa'

It should be noted that systematic orthography does not always provide a sole fixed spelling of a given word. For instance, the above constraint on the well-formed spelling of [ɑ] in German still leaves <a> (8a) and <ah> (8b) as two possible representations of the segment. Instead of using these forms interchangeably, the German writing system has standardised conventions stipulating which form should be used on a word-to-word basis (e.g., <a> for [vɑl] 'whale' but not for [vɑl] 'choice'). MT distinguishes such conventions from systematic orthography and refers to them as *conventional orthography* (Neef, 2012, p. 13; 2015, p. 716).

It is also important to add that systematic orthography is characterised as an optional module (Neef, 2015, p. 716). There are two logical possibilities for an alphabetic system to fully function without this module. The first one is that its graphematics implements a strict one-to-one correspondence between characters and phonological segments.[10] The second one is that the writing system allows more than one characters to represent a single segment and yet relies entirely on conventional orthography to determine the correct spellings of individual words. This latter possibility is not explicitly discussed in the MT literature and requires further exploration in the future.

To summarise, MT explains the script-to-language relationship in alphabetic systems by assuming the three modules described in (9a-c).

(9) Three modules of MT
    a. Language system   provides a linguistic foundation
    b. Graphematics   associates characters with segments
    c. Systematic orthography   optionally decides correct spellings

---

10. As an example of this possibility, Neef (2015, pp. 714–715) cites the International Phonetic Alphabet (IPA) as used for transcribing English. While the IPA can be aptly characterised as a phonologically based alphabetic system (e.g., Coulmas, 2003, pp. 28–33), it is a purpose-built 'technography' as opposed to naturally developed 'orthography' (Mountford, 1996, pp. 627–629). It is open to question whether writing systems belonging to such different categories can be compared on the same level.

2.   Japanese Writing System

Using the key concepts of MT outlined above, this section examines their compatibility with the analysis of the current Japanese writing system. After a brief outline of the underlying language system (Section 2.1), a partial analysis of the writing system is developed in the light of the notions of graphematics (Section 2.2) and systematic orthography (Section 2.3).

2.1.   Language System

Starting with segmental phonology, there has never been a general consensus on how to phonemicise the sounds of modern Japanese. However, assuming the non-derivational linguistic approach of MT, it is possible to make some meaningful generalisations about the sound system at the surface phonological level (based on Vance, 2008, pp. 53–112, 225–232; Labrune, 2012, pp. 25–101, 132–141; Saitō, 2013, pp. 84–96). With regard to vowels, Japanese has five contrastive sounds (10a), each also contrasting with a quantitatively longer counterpart (10b). As for consonants, there are some twenty plain contrastive sounds (10c). Among these, [ɴ] occurs only syllable-finally and is in complementary distribution with [m], [n] and other nasal sounds in that position (e.g., [aɴ], [am.ma], [an.na]). Some consonants have palatalised counterparts (10d), which are not allowed before [e], allophonic before [i], and contrastive before [a o ɯ].[11] Each voiceless obstruent consonant, except for the non-sibilant [ɸ ç h], contrasts with a quantitatively longer counterpart (10e), which occurs only ambisyllabically (e.g., [apːa] = [apˀ.pa]).

(10)   Japanese sound system
     a.   [i e a o ɯ]
     b.   [iː eː aː oː ɯː]
     c.   [p b t d k g ɸ s ɕ ç h ts dz tɕ dʑ m n ɴ ɾ j w]
     d.   [pʲ bʲ kʲ gʲ mʲ ɲ ɾʲ]
     e.   [pː pʲː tː kː kʲː sː çː tsː tɕː]

Moving on to prosodic phonology (based on Vance, 2008, pp. 115–126, 225–232; Labrune, 2012, pp. 142–161; Saitō, 2013, pp. 97–103, 113–116), the above segments are organised into maximally $(C_1)V(C_2)$ syllables (11a-d). The C_2 is either [ɴ] (11c) or an ambisyllabic long consonant occupying the coda position (11d). Japanese is a mora-timed and weight-by-position language (Hayes, 1989, pp. 258–260), wherein a

---

11. While palatalised [tʲ dʲ ɸʲ] have a somewhat similar distribution, they are contrastive only before [ɯ] in a limited number of loanwords, and are often replaced by [tɕ dʑ ç], respectively (e.g., [tʲɯ↓ːba] ∼ [tɕɯ↓ːba] 'tuba').

light (C)V syllable (11a) counts as one mora and a heavy (C)Vː or (C)VC syllable (11b-d) counts as two morae (e.g., Kubozono, 1999, pp. 48−55). Japanese phonology also has lexical pitch accent, which assigns a steep pitch fall to some words (11e) but not to others (11f).[12]

(11)   Japanese syllable structure and lexical pitch accent

     a.   [ka]

     b.   [kaː]

     c.   [kaɴ]

     d.   [kap(ːa)]

     e.   [ka$^{\downarrow}$tːa] 'win-PAST.AFF'

     f.   [katːa] 'buy-PAST.AFF'

Turning now to morphology and word formation (based on Shibatani, 1990, pp. 215−256; Tsujimura, 2013, pp. 125−157; Nitta et al., 2010, pp. 73−92, 225−232), Japanese morphemes may be free or bound, the latter being prefixes, suffixes or enclitics. These morphemes form both monomorphemic (12a) and polymorphemic words, the latter including compounds (12b), derivatives (12c) and inflected items (12d).

(12)   Japanese morphology and word formation

     a.   [ɯta$^{\downarrow}$] 'song'

     b.   [ɯtagoe] 'song+voice' (= singing voice)

     c.   [oɯta] 'HON-song'

     d.   [ɯtatːa] 'sing-PAST.AFF'

Syntactically (based on Shibatani, 1990, pp. 257−262; Tsujimura, 2013, pp. 229−254), Japanese is characterised as a head-final language. It has basic subject-object-verb (SOV) word order, with postpositional particles marking grammatical relations (13).

(13)   Japanese syntax

gakɯseː ga   ɯta$^{\downarrow}$ o   ɯtatːa
student   NOM song ACC sing-PAST.AFF

'A student sang a song.'

With respect to the lexicon (based on Shibatani, 1990, pp. 140−157; Tsujimura, 2013, pp. 229−254; Kageyama and Saito, 2016, pp. 12−29), Japanese lexical items can be classified into four main groups according to their etymological origins. These are known as Native Japanese (NJ) (14a), Sino-Japanese (SJ) (14b), Mimetic (14c) and Foreign (14d). The fifth group of hybrid is also called for because Japanese words include compounds of morphemes from different sublexicons (14e).

---

12. This paper uses a downward-pointing arrow to indicate the position of a pitch fall (Vance, 2008, p. 143).

(14) Japanese lexicon
- a. [kotoba↓] 'word'
- b. [ɡo] 'word'
- c. [wa↓ːdo] 'word' (< Eng. *word*)
- d. [pe↓ɾapera] 'fluent, chitchatty'
- e. [keɴsakuɯwa↓ːdo] 'search word' (SJ + Foreign)

2.2. Graphematics

As noted above, the current Japanese writing system employs a mixture of multiple scripts, namely logographic kanji, moraic hiragana and katakana, and alphabetic rōmaji (Backhouse, 1984, p. 219; Smith, 1996, p. 214; Joyce, 2011, p. 62; Honda, 2012, pp. 38–39).[13] While it is theoretically possible to write Japanese entirely in one script or another, the current norm is to use them all for different purposes in a complementary manner (15).[14] Thus, one may speak of a complex system of written signs divided into typologically distinct but functionally interlinked subparts.

(15) Japanese scripts[15]

Name	Characters represent	Script used to write
Kanji	Morphemes(?)	Lexical items (SJ & NJ)
Hiragana	Morae	Affixes & enclitics (NJ)
Katakana	Morae	Lexical items (Foreign & Mimetic)
Rōmaji	Vowels & consonants	Lexical items (Foreign)

Consequently, Japanese is usually written in a multi-script text (16a), where the characters of different scripts correspond to different properties of linguistic representation (16b).

---

13. Japanese braille (点字 *tenji*) constitutes a separate tactile writing system, which has different formal and functional features from the multi-script visual writing system under discussion (Unger, 1984, p. 254; Hosokawa, 2001, pp. 652–655).

14. Some might find it impossible to write Japanese solely in the kanji script, assuming that they cannot indicate grammatical information. However, this is a viable option given the historical use of *man'yōgana* (万葉仮名) or phonographically used kanji characters (e.g., Seeley, 2000, p. 190).

15. Some notes are in order here. Firstly, the question mark has been added after 'Morphemes' because of uncertainty surrounding the morphemic or morphographic nature of kanji (see below and Section 3). Secondly, hiragana and katakana characters are described as being moraic by some (e.g., Honda, 2012, pp. 72–93) and as core syllabic by others (e.g., Buckley, 2018, pp. 38–42). Thirdly, some scholars use 'arufabetto' (アルファベット) to refer to the Roman script and reserves 'rōmaji' for the Romanised notation of Japanese words (e.g., Satake, 2005, pp. 34–36). Finally, rōmaji (or arufabetto) characters are commonly used for abbreviations of words across sublexicons (ibid., p. 36), in which case they exhibit the characteristics of both phonographs and logographs (Sven Osterkamp and Yannis Haralambous, personal communication).

(16)   Japanese multi-script text

    a. 東京はローマ字で〈Tōkyō〉だ。

    b. toːkʲoː wa ɾoːmaʣi de toːkʲoː da
       Tokyo TOP Rōmaji INS Tokyo COP

    'Tokyo is <Tōkyō> in Rōmaji.'

To account for this fact using the MT model, the present paper proposes to introduce the notion of *structured graphematics*. That is, the graphematic module needs to be divided into submodules, which define the associations between characters or character combinations and linguistic properties in the respective scripts. This is illustrated in (17), showing the graphematic rules for each character used in the example (16) above. The rules are tentative ones for kanji (17a), where it is also possible to interpret the characters as being associated with morphemes (e.g., <東> → {toː} ∨ {higaɕi}; Section 3).

(17)   Japanese graphematics[16]

    a. i. <東> → [toː] ∨ [higaɕi]

        ii. <京> → [kʲoː] ∨ [keː]

        iii. <字> → [ʥi] ∨ [aʣa]

    b. i. <は> → [ha] ∨ [wa]

        ii. <で> → [de]

        iii. <だ> → [da]

    c. i. <ロ> → [ɾo]

        ii. <ー> → [ː]

        iii. <マ> → [ma]

    d. i. <t> → [t]

        ii. <ō> → [oː]

        iii. <k> → [k]

        iv. <y> → [j]

Structured graphematics sufficiently captures both the integrity of the four scripts and their functional division in the Japanese writing

---

16. Many-to-one associations between characters and their linguistic properties are prevalent in the kanji submodule, creating a significant amount of ambiguity in character decoding (Honda, 2012, pp. 156–160). Neef and Balestra (2011, pp. 113–129) use the term *graphematic transparency* to refer to a similar kind of ambiguity in alphabetic systems and propose a way to measure it in German and Italian. Whether their measurement framework is adaptable to Japanese and other non-alphabetic systems remains a topic for future research.

system.[17] Importantly, the kind of correspondence rules assumed for alphabetic systems can also be used for the non-alphabetic kanji, hiragana and katakana (17a–c) as well as the alphabetic rōmaji (17d). Based on this observation, the present paper suggests that assuming the graphematic module is meaningful for the analysis of typologically different writing systems.

2.3. Systematic Orthography

As observed in the German examples discussed above (Section 1.3), the presence of multiple characters representing the same phonological segment generates a graphematic solution space for words containing that segment. According to MT, systematic orthography filters all theoretically possible spellings of a given word and determines the well-formed spelling. If this still leaves more than one spellings, conventional orthography decides the correct one on a word-to-word basis.

   Similar examples of many-to-one correspondence can be found in the graphematic module of the current Japanese writing system. Perhaps unsurprisingly, it is possible to speak of a graphematic solution space with respect to the non-alphabetic submodules (18a–c) as well as the alphabetic one (18d). Thus, regardless of the type of writing system, having multiple ways to represent the same linguistic property necessarily entails multiple ways to write words containing that property.

(18)   Japanese graphematic solution space and orthographic forms[18]
    a.   i.   <見> ∨ <観> → [mi]
        ii.   <見た> [mi↓ta] 'see-PAST.AFF'
        iii.   <観た> [mi↓ta] 'see-PAST.AFF'
    b.   i.   <お> ∨ <を> → [o]
        ii.   <かお> [kao] 'face'
        iii.   <かを> [kao] 'mosquito ACC'
    c.   i.   <ー> ∨ <エ> → [ː] immediately after (C)e

---

17. The same notion may also be used to account for other writing systems. Although the mixed use of multiple scripts is a distinctive characteristic of Japanese, variants of multi-script writing are also found in many writing systems of the world. For example, English employs not only the Roman script but also sets of logographs, ciphers and punctuation marks (see the examples in (2a–d) above). While it is possible to characterise the former as the script of primary importance and the latter of secondary importance, further discussion is needed to establish criteria for such a distinction.

18. In each set of examples presented here, the first line shows two characters or character combinations (separated by the disjunction symbol) that correspond to a single phonological form. The second line presents an example written word containing the more frequent representation of the form in question. The third line gives another example with a less frequent representation.

      ii.   <バレー> [ba↓ɾeː] 'volleyball'

      iii. <バレエ> [ba↓ɾeː] 'ballet'

  d.   i.   <ī> ∨ <ii> → [iː]

      ii. <Īda> [i↓ːda] 'Polish personal name'

      iii. <Iida> [i↓ːda] 'Japanese city name'

Crucially, however, it is difficult to find evidence for a systematic orthography in the Japanese writing system. To illustrate with the kanji examples above, there is no reason to believe that the items in (18a.ii) and (18a.iii) are separate words belonging to different layers of the Japanese vocabulary. In other words, the choice between the two kanji characters is only explicable in terms of conventional rather than systematic orthography. Regarding the hiragana examples in (18b.ii) and (18b.iii), it is conventional to use <お> for [o] in any word and <を> for the accusative particle with the same phonological form.[19] The katakana examples in (18c.ii) and (18c.iii) show the default use of <ー> for representing [ː] and the exceptional word-specific use of <エ> for vowel length in the second syllable.[20] The Rōmaji examples in (18d.ii) and (18d.iii) show proper nouns written in the common Hebonshiki romanisation system; whereas the first character in the former can go with or without the macron (i.e., either <Īda> or <Ida>), the double-character spelling in the latter seems to be the norm for writing the city name (i.e., <Iida> but neither *<Īda> nor *<Ida>).[21]

Nevertheless, it may still be possible to argue for the presence of a systematic orthography in Japanese. It will be recalled that there is a functional division between the four scripts employed in the writing system. This is loosely defined by a set of orthographic guidelines promulgated by the Japanese Cabinet, and is implemented more or less systematically in administration, education and publication. At the same

---

19. While the current norm is to write the former in kanji 顔 and the latter in kanji and hiragana 蚊を, it is possible to write them entirely in hiragana as shown here (see Section 2.2 above and the discussion below for the fungible use of scripts in Japanese). Historically, <お> and <を> were used to represent [o] and [wo], respectively. After the loss of syllable-initial [w] before [o i e] in around 1000 CE (Frellesvig, 2010, pp. 206–207), both characters correspond to [o]. The conventional use of <を> for the accusative particle was codified in the first official guidelines for hiragana orthography promulgated by the Japanese Cabinet in 1946 (現代かなづかい *Gendai Kanazukai* 'Modern Kana Usage'). The same convention is stipulated by the current version updated in 1986 (現代仮名遣い *Gendai Kanazukai* 'Modern Kana Usage').

20. The same syllable is usually written as <レー> as exemplified by <レーザー> [ɾeːdza:] 'laser' and <ジレー> [dʑi↓ɾeː] 'gilet'. Nothing suggests a systematically differentiated representation of [eː] in loanwords based on the source language (e.g., English versus French).

21. The English version of the official website of Iida City in Nagano Prefecture consistently uses this spelling (`https://www.city.iida.lg.jp/`).

time, it is also a common practice to use different scripts interchange-
ably to write the same lexical item for various purposes (Joyce and Ma-
suda, 2019, pp. 255–274). Taking these facts together, one may speak of
a variant of systematic orthography that permits, rather than prohibits,
the fungible use of scripts in this writing system. Further research is
needed to elaborate on the systematicity and flexibility of orthographic
conventions in Japanese and other writing systems.

3.   Logography and Logographic Systems

While MT says very little about non-alphabetic systems, it makes the
following remark concerning logography and logographic systems:

> Logographic writing systems differ from phonographic writing systems
> in that their basic units are logograms, […], i.e., functional classes that corre-
> spond to words or morphemes. (Neef, 2015, p. 713).

As previously mentioned, this statement reflects the traditional clas-
sification of writing systems into two broad types according to the pri-
mary type of written signs (Section 1.2). In the literature, however,
there is an ongoing debate over the validity of this dichotomous divi-
sion. Some studies view all writing systems as being primarily phono-
graphic, even if they may also employ a smaller or larger number of
logographs (e.g., DeFrancis, 1989, pp. 56–64). Some others agree on
the primary importance of phonographs but maintain that logographs
also play an important role in virtually all writing systems (e.g., Sproat,
2000, pp. 139–143). Yet another group of studies reject the primacy
of phonography and classify all writing systems into the phonographic
type and the *morphographic* type (e.g., Joyce, 2011, pp. 63–72).[22] With
respect to MT, the disagreement over the dichotomy between phono-
graphic and logographic systems calls into question the position of lo-
gography in its theoretical framework.

   To address the above question, it would be meaningful to take a closer
look at the graphematic aspects of Japanese kanji, which are widely re-
garded as the prime example of logography (e.g., Sampson, 2015, p. 208;
Sproat, 2000, p. 154). One notable feature of this submodule is that
characters and character combinations correspond to meaning-carrying
phonological forms in many words. For instance, <愛> represents [a↓i],

---

22. Whereas 'logographic / logography' implies the use of mono- and polymor-
phemic word signs, 'morphographic / morphography' suggests that the writing sys-
tem in question primarily employs the signs of free and bound morphemes (Joyce,
2011, pp. 69–70). This latter term is becoming increasingly accepted as an alterna-
tive to the more traditional 'morphemic' label (Rogers, 2005, pp. 14–15; cf. Hill, 1967,
p. 93).

which can form a morphologically simplex word (19a) or an element of complex words (19b,c). A comparison of these and other related words reveals that this phonological form is associated with the meaning 'love'. Because the sound-meaning unit is not analysable into smaller parts, it can be regarded as a morpheme or the minimal meaningful unit in a language. Through a similar analysis, one can say that kanji characters represent morphemes in a large number of kanji-written words (Joyce, 2011, p. 71). This observation appears to support the notion that logographic systems are essentially morphemic and hence 'morphographic' in nature (Section 1.2).

(19)   Morphographic kanji
   a.   <愛> [a↓i] 'love'
   b.   <愛情> [aidʑoː] 'love+emotion' (= affection)
   c.   <恋愛> [ɾeɴai] 'yearning+love' (= romance)

However, it is also important to note that kanji characters do not always represent morphemes. For one thing, they may correspond to phonological forms carrying no discernible meaning. For example, <陛> is only used in the word shown in (20a), where it corresponds to [heː].[23] Although this phonological form historically denoted 'a flight of steps in a palace', such a meaning is synchronically unidentifiable because the character's idiosyncratic usage makes a comparative analysis impossible. A similar description holds for <祉> (20b) as well as for <挨> and <拶> (20c).[24] It is difficult, if not at all impossible, to say that these

---

23. The largest Japanese dictionary *Nihon Kokugo Daijiten* (Kitahara et al., 2000) includes two more headwords containing this character, namely <陛戟> [heːgeki] 'imperial guard' and <楓陛> [ɸɯːheː] 'flight of steps in a palace'. However, they are extremely infrequent in contemporary Japanese, and no instance of either item is found in the 100-million word Balanced Corpus of Contemporary Written Japanese (BCCWJ: https://pj.ninjal.ac.jp/corpus_center/bccwj/en/).

24. A related example can be found in <葡萄> [bɯdoː] 'grape'. Historically, the two constituent characters were invented for the specific purpose of writing the disyllabic monomorphemic word in Chinese. Kanji-written words of this kind can be found in both Chinese and Japanese. A notable feature of these items is the presence of a shared semantic component or 'radical' in both constituent characters (e.g., the three-stroke 'grass' component in <葡> and <萄>), presumably denoting morphological and semantic unity of the word in question (Sproat, 2000, pp. 148–154). Cornelia Schindelin has suggested the term *radical harmony* for this device, in analogy to 'vowel harmony' in phonology (personal communication). In graphematic terms, there are two possible interpretations of radical harmony. The first interpretation is that the constituent characters function as word-specific syllabographs corresponding respectively to the first and second syllables (e.g., <葡> → [bɯ] + <萄> → [doː] when used for {ʙᴜᴅō}; Honda, 2019, p. 202). The second possibility is that they form a digraphic morphograph representing the morpheme in a holistic fashion (e.g., <葡萄> → {ʙᴜᴅō}; Zev Handel and Gordian Schreiber, personal communication). Further discussion is needed to elaborate on this issue.

characters represent morphemes as minimal meaningful units (Honda, 2019, p. 197).

(20) Non-morphographic kanji
　　　a. <陛下> [he↓ːka] 'Majesty'
　　　b. <福祉> [ɸɯkɯ↓ɕi] 'welfare'
　　　c. <挨拶> [a↓isatsɯ] 'greeting'

For another thing, kanji characters may be used phonographically to write certain words, even if they represent individual morphemes elsewhere. This rebus-like use, known traditionally as *ateji* (当て字), is found in many orthographic kanji-written words (21a). The same method is also widely used to produce non-orthographic alternatives to orthographically written words for a more playful stylistic effect (21b).[25] In both cases, individual characters correspond to different portions of the word's phonological form, with little or no regard to their associated meanings. When kanji characters are used this way, there is no reason to assume that they represent individual morphemes (Joyce, 2011, p. 71).

(21) Orthographic and non-orthographic ateji[26]
　　　a.　i.   <出> → [de(↓ɾɯ)] 'go out'
　　　　　ii. <鱈> → [ta↓ɾa] 'cod'
　　　　　iii. <目> → [me↓] 'eye'
　　　　　iv. <出鱈目> [detaɾame] 'hogwash'
　　　b.　i.   <羅> → [ɾa] 'silk gauze'
　　　　　ii. <武> → [bɯ] 'military affairs'
　　　　　iii. <羅武> [ɾa↓bɯ] 'love' (conventionally, katakana <ラブ>)

It should also be added that it is not always clear whether individual kanji characters correspond to morphemes in words with an apparently complex morphological structure (Vance, 2002, p. 187; Honda, 2019, pp. 195–197). To give an example, <勉強> [beɴkʲoː] 'study' is etymologically a compound of <勉> [beɴ] 'strive, serve, fill a post, etc.' and <強> [kʲoː] 'strength, might, strong person, etc.'.[27] From a strictly synchronic standpoint, however, there is little evidence indicating whether or not

---

25. Phonographic use of logographs, as well as logographic use of phonographs, is widely attested across writing systems; see descriptions of individual systems in Daniels and Bright (1996) and Kōno, Chino, and Nishida (2001). This point calls into question the validity of the traditional dichotomy between phonography and logography or morphography (Osterkamp and Schreiber, 2021; cf. Handel, 2020).

26. The characters <出> (21a.i), <目> (21a.iii) and <武> (21b.ii) are also associated with other sound-meaning units, which are omitted here for clarity.

27. In this word, the uvular nasal [ɴ] is phonetically realised as the velar nasal [ŋ] due to anticipatory assimilation of place of articulation. This detail is omitted in the surface phonological transcription adopted in this paper.

the same word retains such compositionality in present-day Japanese. The same can be said for many kanji-written words, both common and uncommon. This observation allows three possible interpretations for the constituent characters in such items: separate morphographs (22a), one polygraphic morphograph (22b) or separate syllabographs with lexically conditioned distributions (22c).

(22) Possible interpretations of <勉強>

    a. <勉> → {beɴ} + <強> → {kʲoː}

    b. <勉強> → {beɴkʲoː}

    c. <勉> → [beɴ] + <強> → [kʲoː] when used for {beɴkʲoː}

The plausibility of the first interpretation depends on specific assumptions about morphological structure and morphemehood (Joyce, 2011, pp. 69–73). The second interpretation requires a theory of what count as polygraphs in different types of writing systems, which is still at an early stage of development (Osterkamp and Schreiber, 2019).[28] The third one implies that each constituent character represents the phonological exponent of the whole or a portion of a morpheme, a claim that needs further examination (Honda, 2019, pp. 202–203).

With respect to MT, the above observations raise several questions about its treatment of logography and logographic systems (23).

(23) Questions about logography and logographic systems

    a. How should MT conceptualise logography in relation to morphography and phonography? Are they mutually exclusive concepts, or do they have commonalities as well as differences?

    b. Is it appropriate for MT to assume the traditional dichotomy between phonographic and logographic (or morphographic) systems as fundamentally different types of writing systems?

    c. Does MT need to make reference to morphology as well as phonology to account for the script-to-language relationship in logographic (or morphographic) systems?

    d. If reference to morphology is necessary, what theories of morphological structure and morphemehood would be compatible with the general framework of MT?

---

28. The term 'polygraph' is usually reserved for the multi-character representation of a single segment or syllable in phonographic systems (e.g., Sproat, 2000, 140, fn. 2). The Japanese writing system employs a considerable number of *jukuji* (熟字) or monomorphemic kanji character combinations, giving rise to the hitherto underexplored notion of 'polygraphic morphographs' or 'morphographic polygraphs' (Honda, 2012, pp. 120–123; Osterkamp and Schreiber, 2019).

4. Conclusion

This paper has explored the adaptability of MT to the analysis of non-alphabetic systems through a discussion of the current Japanese writing system. Using the key concepts of MT, a partial analysis of kanji, hiragana, katakana and rōmaji has been presented to highlight the possibilities and challenges of this theory. A summary of the discussion is given below (24).

(24)   Summary and remaining issues
   a. Current MT is essentially a theory of alphabetic systems. However, with its non-derivational linguistic approach, MT has the potential for expanding into a general theory of script-to-language relationship across different types of writing systems.
   b. In principle, the theory's three-module model is adaptable to the analysis of the Japanese writing system. This suggests that the key concepts of MT are applicable to non-alphabetic systems.
   c. However, some of the basic assumptions about graphematics and systematic orthography require modification in this context:
      i. The notion of structured graphematics should be introduced to account for the functional division between kanji, hiragana, katakana and rōmaji.
      ii. Further research is needed to elaborate on the systematicity and flexibility of orthographic conventions in view of the choice of characters and the fungible use of scripts in Japanese.
   d. Observations on the graphematic aspects of kanji characters call for further discussion of issues related to logography and logographic systems.

In conclusion, a MT approach to the Japanese writing system provides a new perspective on the capability of this theory to account for the script-to-language relationship in different types of writing systems. Further research is called for to elaborate and examine the generality of MT.

References

Backhouse, Anthony E. (1984). "Aspects of the graphological structure of Japanese." In: *Visible Language* 18.3, pp. 219–228.

Bloomfield, Leonard (1933). *Language*. London: George Allen and Unwin.

Buckley, Eugene (2018). "Core syllables vs. moraic writing." In: *Written Language and Literacy* 21.1, pp. 26–51.

Chomsky, Carol (1970). "Reading, writing, and phonology." In: *Harvard Educational Review* 40.2, pp. 287–309.

Chomsky, Noam and Morris Halle (1968). *The sound pattern of English*. New York: Harper and Row New York.

Coulmas, Florian (2003). *Writing systems: An introduction to their linguistic analysis*. Cambridge: Cambridge University Press.

Daniels, Peter T. (1996). "The study of writing systems." In: *The world's writing systems*. Ed. by Peter T. Daniels and William Bright. New York: Oxford University Press New York, pp. 3–17.

——— (2018). *An exploration of writing*. Sheffield and Bristol: Equinox Publishing Limited.

Daniels, Peter T. and William Bright (1996). *The world's writing systems*. New York: Oxford University Press.

DeFrancis, John (1989). *Visible speech: The diverse oneness of writing systems*. Honolulu: University of Hawaii Press.

Douglas, Aileen (2017). *Work in hand: Script, print, and writing, 1690-1840*. Oxford: Oxford University Press.

Faber, Alice (1992). "Phonemic segmentation as epiphenomenon: Evidence from the history of alphabetic writing." In: *The Linguistics of Literacy* 21, pp. 111–134.

Frellesvig, Bjarke (2010). *A history of the Japanese language*. Cambridge: Cambridge University Press.

Gelb, Ignace J. (1963). *A study of writing*. Chicago and London: University of Chicago Press.

Gnanadesikan, Amalia E. (2017). "Towards a typology of phonemic scripts." In: *Writing Systems Research* 9.1, pp. 14–35.

Haas, William (1983). "Determining the level of a script." In: *Writing in focus*. Ed. by Florian Coulmas and Konrad Ehlich. Berlin: Mouton, pp. 15–29.

Handel, Zev (July 18, 2020). *Is Logographic a Valid Script Category? Evidence from Historical Borrowing of the Chinese-Character Script*. Grapholinguistics in the 21st Century–From Graphemes to Knowledge. URL: https://youtu.be/Z1Usj8JqruA (visited on 12/16/2020).

Haralambous, Yannis (2020). "Grapholinguistics, TₑX, and a June 2020 conference." In: *TUGboat* 41.1, pp. 12–19.

Harris, Roy (1995). *Signs of writing*. London: Routledge.

Hayes, Bruce (1989). "Compensatory lengthening in moraic phonology." In: *Linguistic Inquiry* 20.2, pp. 253–306.

Hill, Archibald A. (1967). "The typology of writing systems." In: *Papers in linguistics in honor of Leon Dostert*. Ed. by William M. Austen. The Hague and Paris: Mouton The Hague, pp. 92–99.

Honda, Keisuke (2012). "The relation of orthographic units to linguistic units in the Japanese writing system: An analysis of kanji, kana and kanji-okurigana writing." PhD thesis. University of Tsukuba.

———— (2019). "What do kanji graphs represent in the current Japanese writing system? Towards a unified model of kanji as written signs." In: *Proceedings of Graphemics in the 21st Century, Brest 2018*. Ed. by Yannis Haralambous. Brest: Fluxus Editions, pp. 185–208.

Hosokawa, Yukiko [細川由起子] (2001). "点字 [Braille]." In: 言語学大辞典別巻世界文字辞典 *[The Sanseidō encyclopaedia of linguistics, Vol. 7: Scripts and writing systems of the world]*. Ed. by Rokurō Kōno [河野六郎], Eiichi Chino [千野栄一], and Tatsuo Nishida [西田龍雄]. 東京 [Tokyo]: 三省堂 [Sanseido Press], pp. 641–655.

Joyce, Terry (2011). "The significance of the morphographic principle for the classification of writing systems." In: *Written Language and Literacy* 14.1, pp. 58–81.

———— (2016). "Writing systems and scripts." In: *Verbal communication*. Ed. by Andrea Rocci and Louis De Saussure. Berlin and Boston: De Gruyter Mouton, pp. 287–308.

Joyce, Terry and Susanne R. Borgwaldt (2011). "Typology of writing systems: Special issue introduction." In: *Written Language and Literacy* 14.1, pp. 1–11.

Joyce, Terry and Hisashi Masuda (2019). "On the notions of graphematic representation and orthography from the perspective of the Japanese writing system." In: *Written Language and Literacy* 22.2, pp. 247–279.

Justeson, John S (1976). "Universals of language and universals of writing." In: *Linguistic studies offered to Joseph Greenberg on the occasion of his sixtieth birthday*. Ed. by Alphonse Juilland. Saratoga: Anma Libri, pp. 57–94.

Kageyama, Taro and Michiaki Saito (2016). "Vocabulary strata and word formation processes." In: *Handbook of Japanese lexicon and word formation*. Ed. by Taro Kageyama and Hideki Kishimoto. Boston and Berlin: Walter de Gruyter Inc., pp. 11–50.

Kitahara, Yasuo [北原保雄] et al., eds. (2000). 日本国語大辞典 *[Great dictionary of the Japanese language]*. 2nd ed. 東京 [Tokyo]: 小学館 [Shōgakukan].

Kōno, Rokurō [河野六郎], Eiichi Chino [千野栄一], and Tatsuo Nishida [西田龍雄] (2001). 言語学大辞典別巻世界文字辞典 *[The Sanseidō encyclopaedia of linguistics, Volume 7: Scripts and writing systems of the world]*. 東京 [Tokyo]: 三省堂 [Sanseidō].

Kubozono, Haruo (1999). "Mora and syllable." In: *The handbook of Japanese linguistics*. Ed. by Natsuko Tsujimura. Malden: Blackwell Publishers, pp. 31–61.

Labrune, Laurence (2012). *The phonology of Japanese*. Oxford: Oxford University Press.

Meletis, Dimitrios (2019). "The grapheme as a universal basic unit of writing." In: *Writing Systems Research* 11.1, pp. 26–49.

Mountford, John (1996). "A functional classification." In: *The world's writing systems*. Ed. by Peter T. Daniels and William Bright. New York: Oxford University Press, pp. 627–632.

Neef, Martin (2012). "Graphematics as part of a modular theory of phonographic writing systems." In: *Writing Systems Research* 4.2, pp. 214–228.

————— (2015). "Writing systems as modular objects: Proposals for theory design in grapholinguistics." In: *Open Linguistics* 1, pp. 708–721.

Neef, Martin and Miriam Balestra (2011). "Measuring graphematic transparency: German and Italian compared." In: *Written Language and Literacy* 14.1, pp. 109–142.

Nitta, Yoshio [仁田義雄] et al. (2010). 現代日本語文法 *[Modern Japanese grammar]*. 東京 [Tokyo]: くろしお出版 [Kurosio Shuppan].

Osterkamp, Sven and Gordian Schreiber (Mar. 26, 2019). *<Th>e ubi<qu>ity of polygra<ph>y and its significance for <th>e typology of <wr>iti<ng> systems*. The 12th International Workshop of the Association for Written Language and Literacy. URL: http://faculty-sgs.tama.ac.jp/terry/awll/wsMaterials/AWLL12-2019-ProgrammeAbstracts.pdf (visited on 12/16/2020).

————— (2021). "Challenging the Dichotomy Between Phonography and Morphography: Transitions and Gray Areas." this volume.

Pettersson, John Sören (1996). "Numerical notation." In: *The world's writing systems*. Ed. by Peter T. Daniels and William Bright. New York: Oxford University Press, pp. 795–806.

Ratcliffe, Robert R. (2001). "What do 'phonemic' writing systems represent?: Arabic huruuf, Japanese kana, and the moraic principle." In: *Written Language and Literacy* 4.1, pp. 1–14.

Rogers, Henry (2005). *Writing systems: A linguistic approach*. Malden: Blackwell Publishing.

Saitō, Yoshio [斎藤純男] (2013). 日本語音声学入門 *[Introduction to Japanese phonetics]*. 東京 [Tokyo]: 三省堂 [Sanseidō].

Sampson, Geoffrey (2015). *Writing systems*. Sheffield and Bristol: Equinox Publishing Ltd.

Sasahara, Hiroyuki [笹原宏之] (2001). "日本の文字 [Japanese scripts]." In: 言語学大辞典別巻世界文字辞典 *[The Sanseidō encyclopaedia of linguistics, Volume 7: Scripts and writing systems of the world]*. Ed. by Rokurō Kōno [河野六郎], Eiichi Chino [千野栄一], and Tatsuo Nishida [西田龍雄]. 東京 [Tokyo]: 三省堂 [Sanseido Press], pp. 696–706.

Satake, Hideo [佐竹秀雄] (2005). "現代日本語の文字と書記法 [Scripts and writing in modern Japanese]." In: 文字・書記 *[Scripts and writing]*. Ed. by

Hayashi Chikafumi [林史典]. 東京 [Tokyo]: 朝倉書店 [Asakura Shoten], pp. 22–50.

Saussure, Ferdinand de (1983/1916). *Course in general linguistics*. Trans. by Roy Harris. London: Duckworth.

Seeley, Christopher (2000). *A history of writing in Japan*. Honolulu: University of Hawai'i Press.

Shibatani, Masayoshi (1990). *The languages of Japan*. Cambridge: Cambridge University Press.

Smith, Janet S. (Shibamoto) (1996). *Japanese writing*. Ed. by Peter T. Daniels and William Bright. New York: Oxford University Press, pp. 209–217.

Sproat, Richard (2000). *A computational theory of writing systems*. Cambridge: Cambridge University Press.

Taylor, Insup and M. Martin Taylor (2014). *Writing and literacy in Chinese, Korean and Japanese*. Rev. ed. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Tsujimura, Natsuko (2013). *An introduction to Japanese linguistics*. Malden and Oxford: John Wiley & Sons.

Unger, J. Marshall (1984). "Japanese braille." In: *Visible Language* 18.3, pp. 254–256.

Vachek, Josef (1973). *Written language: General problems and problems of English*. The Hague: Mouton.

Vance, Timothy J. (2002). "The exception that proves the rule: Ideography and Japanese kun'yomi." In: *Difficult characters: Interdisciplinary studies of Chinese and Japanese writing*. Ed. by Mary S. Erbaugh. Columbus: National East Asian Language Resource Center, Ohio State University, pp. 177–193.

———— (2008). *The sounds of Japanese*. Cambridge: Cambridge University Press.

Levels of Structure
Within Chinese Character Constituents

James Myers

Abstract. It is already known that Chinese readers and writers decompose characters into four structural levels: basic components, complex strokes, simple strokes, and stroke features. These levels parallel word-internal structure in spoken and signed languages (respectively, morphemes, complex segments, segments, and segmental features). In this paper I consider evidence for a level intermediate between basic components and strokes: the stroke group. Like syllables, stroke groups are targeted by stress-like prominence and analyzable in terms of analogs to onsets, nuclei, and codas. They also seem to compete with each other for space within a component, as syllables do within morphemes. Though the analogies between stroke groups and syllables are weaker than the linguistic analogies for other character levels, the stroke group concept may help improve our understanding of a hitherto understudied aspect of writing systems: stroke interactions.

1. Introduction

Myers (2019) is a sober defense of an outrageous idea, the idea that Chinese characters conform to a genuine lexical grammar. In it I argue that Chinese script has direct analogs to morphemes, affixation, compounding, reduplication, inflectional agreement, idiosyncratic allomorphy, regular allomorphy, prosodic structure, stress, stress clash, weight, distinctive features, and feature spreading. The arguments are backed up with evidence from quantitative corpus analyses and psycholinguistic experiments, and in general I tried to be cautious in advancing only those claims that seemed reasonably well-established empirically.

In this paper I take a somewhat more, shall we say, speculative approach. Namely, while in the book I mused on whether certain types of stroke groups in Chinese characters are analogous to syllables, I did not

James Myers   [ID] 0000-0002-3866-1969
Graduate Institute of Linguistics, National Chung Cheng University, 168 University Road, Min-Hsiung, Chia-Yi 62102, Taiwan
E-mail: Lngmyers@ccu.edu.tw

push the idea. The primary purpose of this paper is to see how far this idea can go anyway.

Before we begin, I should say that, like Myers (2019), this paper focuses on traditional characters, but I do occasionally allude to simplified characters, which have almost exactly the same grammar. I also focus only on structures and patterns in modern characters that are hypothesized to be mentally active in contemporary readers and writers; despite its importance to other aspects of character analysis, etymology is thus irrelevant here.

Chinese characters have long been recognized as having multiple levels of representation. For example, the character in (1a) consists of the complex constituents in (1b-c), synchronically interpretable in terms of meaning and/or pronunciation. The constituent in (1c), in turn, consists of the basic components in (1d-e), where that in (1e) is not synchronically interpretable, but appears in other characters like those in (1f). The component in (1e) can be further decomposed into the strokes in (1g), including simple strokes, like the vertical stroke that forms its left edge, and complex strokes, like the rotated-L shape that forms the upper right corners of its two boxlike substructures.

(1)   a.   館 *guǎn* 'public building'
      b.   食 *shí* 'meal'
      c.   官 *guān* 'government official'
      d.   宀 *mián* (roof-related semantic marker)
      e.   𠯑 (synchronically lacking meaning and pronunciation)
      f.   師 *shī* 'army' 遣 *qiǎn* 'dispatch'
      g.   丨 𠃌 一 𠃌 一

As will be reviewed in section 2, all of these levels have been demonstrated to be mentally active in the minds of modern readers and to have relatively self-evident analogs to levels in the internal structure of spoken and signed words.

Section 3 then explores the proposed level of stroke groups, which lies between the levels of components and strokes, and points out several similarities they share with syllables. For example, the component in (1e) above contains two boxes, each composed of more than one stroke, and joined together rather than being separate components. Since the lower box in (1e) is larger than the upper one, it is being treated as a whole by some sort of enlargement process. This process is argued in Myers (ibid.) to be like stress, making the targeted stroke group analogous to a stressed syllable. The fact that the box is treated as a whole is also consistent with how its strokes interact (i.e., are arranged with respect to each other), and to a large extent, interactions within stroke groups prove to be analyzable in terms of analogs to syllable-internal structure like onsets, nuclei, and codas. Moreover, stroke groups seem

to compete with each for space within components, much as syllables do within morphemes.

Section 4 ends the paper with some conclusions.

2.   Levels of Structure in Chinese Characters

In spoken and signed languages, morphological and phonological structures each consist of hierarchical levels, though the two types of hierarchies themselves parallel each other. For example, in the American English pronunciation of the word in Figure 1, the division between morphemes does not correspond precisely to that between syllables (σ), because the /t/ is ambisyllabic between the strong (stressed) and weak (unstressed) syllables, as defined by the metrical foot $[SW]_F$ (the moras reflect segmental duration and syllable weight; see, e.g., Hayes, 1989). This prosodic structure also causes the /t/, lexically specified with the feature [-voiced], to be realized as [+voiced] [ɾ].

'eating'
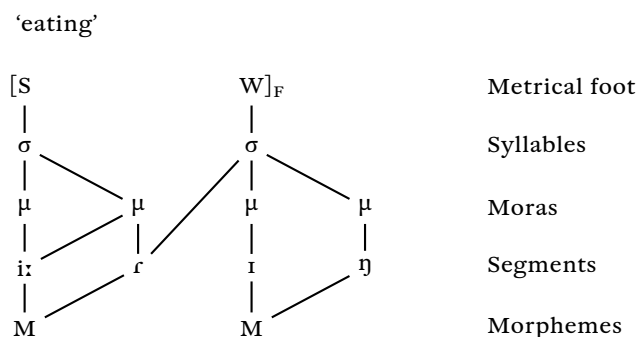
FIGURE 1. Autosegmental analysis of an American English word

The goal of this section is to review evidence that Chinese characters also have a hierarchical structure. Even ancient Chinese linguists recognized that characters are composed of interpretable components, which in turn are built using a small inventory of strokes, but as shown in sections 2.1 and 2.2, these levels have more recently also been thought of as corresponding to morphemes and segments, respectively. In 2.3 I review arguments from Myers (2019) suggesting that both of these levels also interact with something like the metrical structure of spoken and signed phonology.

2.1.   Components

The best-studied level of Chinese character structure is the component. The most transparent of these is the semantic radical, which prototypically relates to the meaning of the whole character, as illustrated in (2a). Characters also often have a so-called phonetic component, which hints at the character's pronunciation, though most of these are also complex constituents containing more than one component, as illustrated in (2b).

(2)   a.   婚 *hūn* 'marry'   女 *nǚ* 'female'
      b.   昏 *hūn* 'dusk'   氏 *shì* (a surname)   日 *rì* 'sun'

All characters can ultimately be decomposed into basic components, forming an inventory much smaller than that of characters. As estimated by Hue (2003), educated traditional character readers know over 5,000 characters, and Unicode contains many tens of thousands, but estimates for the number of basic components in traditional characters ranges only from around 250 to around 650 (see Myers, 2019, Section 1.2.2.3). The component inventory cannot be definitively fixed in part because the character inventory is not fixed, and in part because not all components are interpretable (see, e.g., Slaměníková, 2018). For example, the character in (3a) clearly contains the (uninterpreted) component in (3b), but the rest of the character is not found anywhere else in the traditional character system; the character in (3c) is the simplified equivalent of (3a). Chuang and Teng (2009) treat (3a) as an atomic component, whereas Lu, Chan, Li, and Li (2002), who also cover simplified characters, treat the bottom portion as a component in its own right; Wikimedia Commons[1] instead decomposes it into the components in (3d). For consistency in this paper, I will pretend that the inventory of 441 traditional character components proposed by Chuang and Teng (2009) is definitive.

(3)   a.   單 *dān* 'single'
      b.   口 *kǒu* 'mouth'
      c.   单 *dān* 'single' (simplified system)
      d.   甲 *jiǎ* 'shell'   一 *yī* 'one'

Like morphemes, character components are the minimal potentially interpretable units, even if, like morphemes, not all are actually interpretable synchronically, as in English *result*, *resist*, *consult*, *consist* (see Aronoff, 1994 for the notion of "morphology by itself"). Also like genuine morphology, character decomposition is often recursive, as illustrated above in (2). Such parallels have often been noted (Ladd, 2014;

---

1. https://commons.wikimedia.org/wiki/Commons:Chinese_characters_decomposition.

Feldman and Siok, 1999), but Myers (2019) takes them further, noting, among other things, that the formal and functional properties of semantic radicals have much in common with those of inflectional affixes, and that component reduplication, as in the top of (3a), shares formal and functional properties with its namesake in spoken and signed morphology (see also Behr, 2006).

There is copious evidence that readers and writers mentally activate character components (see Myers, 2019, Chapter 5, for a thorough review). Taft and Zhu (1997), for example, found that characters were recognized more quickly if they contained higher-frequency components, even if they were unrelated in meaning and pronunciation to the character as a whole; Chen and Cherng (2013) drew related conclusions from handwriting experiments. Such observations are consistent with corpus modeling. For example, Li and Zhou (2007) showed that characters share components to the precise degree that one would expect if characters were generated from components via a general grammar rather than via exemplar-driven analogy (see also Fujiwara, Suzuki, and Morioka, 2004; Haralambous, 2013).

One particular sort of corpus-based generalization will prove particularly relevant to the present study, since it was first observed in syllables and other phonological units: Menzerath's law (Menzerath, 1954), also called the Menzerath-Altmann law (Altmann, 1980). Informally, this law states that the more constituents within a constituent at the next-higher level (e.g., syllables within a morpheme), the simpler they tend to be. In other words, the lower-level units compete for space within the higher-level unit.

This law applies to character components as well. As shown for Japanese Kanji by Prün (1994), and for simplified Chinese characters by Bohn (1998), the more components a character has, the fewer their mean number of strokes; independently, Chen and Liu (2019) showed that this generalization also holds for components in a multi-character word (probably, I speculate, because longer Chinese words tend to be transliterations of foreign borrowings, which tend to reuse the same small set of relatively simple characters).

Formally, Menzerath's law is an inverse power law, as in the simplified version in (4) given in Prün (1994, p. 149), whereby the mean size or complexity y of the lower-level constituents is correlated (inversely, given the negative b) with the complexity of the upper-level constituent x (in terms of the number of lower-level constituents), nonlinearly, so that the difference in constituent size for one versus two constituents is larger than that between two versus three, and so on.

(4) $y = ax^b, \ b < 0$

As Prün (ibid.), Bohn (1998), Chen and Liu (2019) have all noted, the fact that the law applies to character components suggests that they rep-

resent a genuine level of description. It thus provides a tool for testing for the potential psychological reality of character levels even without running a psycholinguistic experiment.

2.2. Strokes

All writing is composed of strokes, that is, marks left via continuous contact between writing instrument and writing surface. Modern Chinese script has a set of basic linear strokes that can be used in their basic forms, concatenated with each other (i.e., changing stroke axis without lifting the writing instrument), and/or slightly modified (via curving and/or hooking), as shown in (5) (strokes that fit more than one category are repeated).

(5)   a.   Simple strokes: ` 一 丨 丿 丿 乀 丿 乛 乚 亅 ⺄ 乀 乀 乀

      b.   Complex strokes: 𠃌 𠃊 乚 𠃍 ㄑ ㄥ ㄟ ㄋ 𠃌 乚 ㇋ ㇌ ㇉ ㄅ ㄟ ㄅ ㄋ ㄋ

          ㄅ 乙 ㄋ

      c.   Curving: 乀 丿 丿 丿 乚 乀 丿 ㄟ ㄋ ㄋ ㄋ

      d.   Hooking: 丿 乚 乀 乚 亅 丿 乛 乚 𠃌 乚 ㄅ ㇉ 乙 ㄋ ㄋ

As has often been observed (Wang, 1983; Peng, 2017; Myers, 2019), strokes are like phonological segments in being basic units that are readily analyzable in terms of distinctive features, as opposed to the multi-stroke morpheme-like components that they compose. Watt (1980) observed a similar three-level contrast (morpheme = letter, stroke, feature) in the Roman alphabet.

No matter how Chinese stroke features are formalized, they serve to encode axis (horizontal, vertical, main diagonal [\\], counterdiagonal [/]), curving, and hooking. Though strokes are visual marks, they also require encoding in terms of motoric gestures, much as evidence for character components comes from both perceptual and production experiments (see Myers, 2019, Sections 1.3.1.4 and 5.2.1.2, for more on amodality as a sign of the grammar-like nature of the Chinese character system). Individual stroke direction is mostly from left and/or top to right and/or bottom (for reasons that will be discussed in a later section). This gives the so-called dot, the simplest of all strokes, its default direction along the main (falling) diagonal (see first stroke in (5a) above). Since the counterdiagonal axis cannot be drawn simultaneously left to right and top to bottom, the Chinese stroke inventory offers two distinct strokes: that in (6a) is written top to bottom but right to left, whereas that in (6b) is written from left to right but bottom to top.

(6)   a.   才

      b.   子

The inventory of complex strokes is limited by the same stroke direction constraints: the axis direction changes at the endpoint (right or bottom) of the previous portion, as in (7).

(7)   a.   乛 司
        b.   乛 口
        c.   ∟ 匠

Strokes have undoubted psychological reality. Readers and writers tend to be consciously aware of them because they are explicitly referred to in lexicographical and pedagogical traditions, but they affect automatic processing as well: stroke number is routinely taken into account in reading experiments to control for visual complexity, and there is some evidence that it matters in writing experiments as well (Wang, Huang, Zhou, and Cai, 2020).

As demonstrated by Bohn (1998), Menzerath's law also applies to strokes, further reconfirming their psychologically real status. Stroke complexity was quantified on a scale that counted the number of linear segments and also hooking, so that both strokes in (8a) were given a score of two, on up to a maximum score of 5 for the stroke in (8b) (four segments plus a hook). Bohn found that the more strokes there were in a character component, the lower was the mean stroke complexity, in accordance with an inverse power function.

(8)   a.   乛 亅
        b.   乃

2.3.   The Prosodic Structure of Chinese Characters

Building on research like that sketched above, Myers (2019) argues that parallel to the traditional hierarchy of strokes building components building characters, there is also structure analogous with prosody, specifically metrical feet. In spoken language, metrical feet are supported by a vast array of data, including perception experiments (Cutler and Clifton, 1984), production experiments (Levelt, Roelofs, and Meyer, 1999), and corpus analyses (Myers and Tsay, 2015), and there are similar data from signed languages as well (Crasborn, Kooij, and Ros, 2012).

The notion that written forms have a visual analog to prosody is also advanced in works like Evertz (2018) for spelling in the Roman alphabet. However, while the visual prosody of letters is claimed to correlate with that of the spoken phonemes they represented, the prosody that Myers (2019) sees in Chinese characters is completely unrelated to any of the spoken languages written with them, but relates solely to visual form.

At the heart of the analysis is the prosodic template in (9), realized in its full form as in (9a) and reduced as in (9b-d). This template places a single strong (S) position (head) in the bottom and/or right of a character or a component, with the remaining positions being weak (W). The lower right position is emphasized because strokes within a component and components within a character all tend to be written from left to right and top to bottom, and final gestures are given greater emphasis, not just in writing (e.g., in Western handwriting: Wann and Nimmo-Smith, 1991) but also in speech (Beckman and Edwards, 1990) and signing (Sandler, 1993).

(9)   a. $\begin{bmatrix} W \\ W\ S \end{bmatrix}$

       b. $\begin{bmatrix} W \\ S \end{bmatrix}$

       c. $[W\ S]$
       d. $[S]$

This stress-like prominence is visually obvious in the different sizes of the reduplicated components in (10), where the larger component is at the bottom (10a) or right (10b). Note also that component reduplication involves doubling, either along one axis (10a-b) or along both (10c), again similar to the prosodically constrained reduplication of spoken language (McCarthy and Prince, 1998) and signed language (Berent and Dupuis, 2017).

(10)   a.   昌 多 炎
       b.   林 珏 比
       c.   品 森 犇

This prominence generalization is sometimes clear even when the components are different, as in (11).

(11)   a.   大：奥～奇
       b.   田：富～畢

Myers (2019) argues that the prosodic weakness of the leftmost and topmost positions also explains (synchronically) the preference for semantic radicals to appear in these positions. This is because semantic radicals tend to be small along the relevant dimension (i.e., left-edge radicals are thin relative to the horizontal axis while top-edge radicals are flat relative to the vertical axis), as in (12).

(12)   a.   彳：很
       b.   艹 (derived from 艸)：花

The same correlation between position and size is also observed within basic components, as illustrated in (13) with some from the inventory proposed for traditional characters by Chuang and Teng (2009). Note that among the parallel strokes, the longest is that at the bottom (13a) or right (13b).

(13) a. 二 干 土 彳 彡
      b. 开 川

These generalizations have exceptions. Characters with small right-edge or bottom-edge semantic radicals do exist, like those in (14a), and there are an even smaller number of components with prominent top-most horizontal strokes, as in (14b). The existence of exceptions is also stresslike, however: the strong preference of English for strong-weak stress (*button*) and unstressed suffixes (*eating*) is not nullified by exceptions (*baton*, *unwell*).

(14) a. 戈：戰
      b. 士

Some of the above components also illustrate an analog of prosodically conditioned allophony, namely curving of the vertical stroke on the left edge. Further components showing this are given in (15).

(15) 廾 井 片 用

The curving generalization seems to have more lexical exceptions than prominence, as illustrated with the components in (16).

(16) 巾 冉 冊

However, as first observed by Wang (1983), such exceptions are more likely in horizontally wider constituents, where, for example, the characters in (17a) have more horizontal strokes (making them "taller") than the corresponding ones in (17b) (making them "wider").

(17) a. 冂：周 月
      b. 冂：同 冊

Myers (2019) confirms that this tendency is indeed statistically significant, then goes on to propose an explanation: curving is only possible in a prosodically weak position, and wider components contain two prosodic templates rather than one, putting the left edge in a strong (head) position. This analysis is illustrated in (18). Curving in a weak position is thus similar to vowel reduction in unstressed syllables.

(18)   a.   月 [WS]
       b.   册 [S][S]

In addition to statistical analyses of character databases, the psychological reality of the character prosodic template has also been supported in a series of experiments. Myers (2016) demonstrated that readers generalize constraints on reduplication shape to nonce characters. Myers (2019) showed that readers find nonce components more acceptable if the largest stroke is at the bottom or right, and prefer a curved vertical stroke to appear at the left rather than elsewhere. Myers (2020) found that readers prefer nonce characters to have thin rather than wide left-edge semantic radicals, whereas widening right-edge semantic radicals did not reduce acceptability much, presumably because doing so made the nonce characters conform better to the regular prosodic template.

Note that because metrical feet and the proposed character prosodic template are defined by their shape, Menzerath's law does not apply. It would make no sense to ask if feet are smaller in longer words, because feet always have the same number of syllables (if available), and the same is true for the proposed character template.

3.   Stroke Groups As Orthographic Syllables

As spelled out in (19), spoken and signed syllables have a number of well-established properties.

(19)   a.   In a syllable, sonority (energy) increases to a peak and then falls.
       b.   Syllables are perceptually highly salient.
       c.   Syllables are targeted by foot-level processes like stress.
       d.   Articulatory gestures are more closely coordinated within than across syllables.
       e.   Nuclei are obligatory, onsets are favored, and codas are disfavored.
       f.   Syllables compete for space in morphemes (Menzerath's law).

Properties (19a-b) do not seem to apply to stroke groups. Whereas we can say that /pro/ makes a good syllable and */rpo/ a bad one (and likewise for <pro> vs. *<rpo>, as reviewed in Evertz, 2018) because of the intrinsic sonority of the segments (and letters), there seems to be no way to rank Chinese strokes in an analogous way. Strokes do differ in energy (as reflected in size), but as we saw in the previous section, this is predictable from position, making this phenomenon analogous to stress and not sonority. Moreover, there is as yet no evidence that stroke

groups are perceptually salient, that is, that readers are sensitive not just to components and strokes but also to some intermediate level.

Nevertheless, section 3.1 argues that property (19c) does apply to stroke groups: the character template is built on syllable-like units. Section 3.2 then provides arguments that the related properties (19d-e), concerning syllable-internal structure, apply as well. Finally, Section 3.3 demonstrates property (19f): the applicability of Menzerath's law.

3.1.   Stroke Groups and Prosodic Regularities

If regular prominence at the bottom and right is analogous to stress, enlargeable constituents should be analogous to syllables. As we have seen, these include certain simple components, like those in (20a), and certain individual strokes, as in (20b).

(20)   a.   昌 多
       b.   二 土 川 井

Yet regular enlargement also affects groups of multiple strokes, not just individual ones, even if they do not form full components. I illustrate this in (21) with a variety of examples: (21a-c) show enlargement of the lower of two linked boxes, (21d) shows something similar with other duplicate sets of strokes, and (21e) shows a cross-character contrast in enlargement of box versus (linked) stroke.

(21)   a.   串 弗 虽
       b.   龟
       c.   官
       d.   出 飛
       e.   由 甲

At the same time, not all individual strokes are subject to prominence. As illustrated in (22), prominence does not affect the bottommost horizontal stroke if this ends (at the right) at another stroke (22a) or makes contact at both ends (22b), and instead the next-lowest free horizontal stroke is enlarged. If there is no free stroke, as in (22c), prominence can only apply to the entire complex, as in (22d). Bottommost strokes that cross others but are free at both ends (22e), and perhaps also those free just at the right (22f), are subject to prominence, as are strokes that are contacted at their midpoint by other strokes, as in (22g).

(22)   a.   尹 (𠃌 一 一 丨)
       b.   廿 (一 丨 丨 一)
       c.   口 (丨 𠃌 一)

    d.   串
    e.   干
    f.   非
    g.   工

One way to put all of these observations together is to consider spa-
tially separated strokes (and certain simple components, to be eluci-
dated later) to be stroke groups, along with simple strokes with free
ends. A stroke that ends in contact with another stroke is instead part
of a stroke group that contains both strokes, unless that stroke is subject
to a stroke-group-level process, like prominence, and thus a separate
stroke group.

Left-edge curving also provides some information about the nature
of stroke groups. According to the argument given in section 2.3, a ver-
tical curved stroke is only possible if it is in a prosodically weak position,
analogous to an unstressed syllable. I also argued that a vertical stroke
in this position is more likely to be straight if it is the head of its own
prosodic template, analogous to a stressed syllable. Either way, then, a
potentially curvable leftmost vertical stroke should be considered a sep-
arate stroke group. Thus despite being composed of contacting strokes,
each of the components in (23) should contain at least two stroke groups.

(23)   厂 尸 冂

3.2.   Stroke Groups and Stroke Interactions

If stroke groups are like syllables, they should also restrict how strokes
can combine, similar to the way spoken and signed syllables restrict
phoneme and handshape sequences. I start my argument for this claim
in section 3.2.1 with a review of previous studies on stroke interactions
in the perception and production of simple line drawings, and then show
how these relate to the structure of syllables in speech and signing.
Combined with the previous discussion of the prosody of prominence
and curving, this comparison will allow me in section 3.2.2 to interpret
different kinds of basic stroke interactions in terms of different kinds of
syllable-internal structure. Particularly challenging cases are surveyed
in section 3.2.3.

3.2.1.   Natural Stroke Interactions

A particularly insightful analysis of how perception affects written
strokes is given in Changizi, Zhang, Ye, and Shimojo (2006), who
counted the frequencies of all 36 possible configurations of one to three

strokes in a variety of writing systems and beyond. Each of their configurations defined a class of topological equivalents, where, for example, <Z> and <[> are identical since both link three strokes at two joints. For my purposes, their key finding was that writing systems strongly favor a small subset of configurations, with only those listed in (24) approaching or exceeding a proportional frequency of .1 (taken from Figure 2, p. E118). Each configuration is illustrated with Chinese character components that contain it.

(24) a. I   一二八刂川
      b. L   乚乛厂乀
      c. T   丁卜亻人入
      d. X   乂十
      e. Z   匚凵冂乙𠃌
      f. F   乚

Changizi, Zhang, Ye, and Shimojo (ibid.) argue that the variation in configuration frequency is due to visual and not motoric processes, since the same variation is observed in trademarks, which are virtually never handwritten, but not in shorthand, where writing ease is favored over visual clarity.

Nevertheless, as noted in section 2.2, strokes are also gestural things, having not just an axis but also a direction (i.e., they are vectors), with the strong preference for the rightward and downward directions constraining what complex strokes are possible. Seeing strokes as vectors also helps explain stroke combinations as well. In particular, in Chinese character components, the T configuration is not only quite common, but is almost always written with the midpoint of one stroke (e.g., the top of the T) coinciding with the starting, not the ending, of the other stroke (e.g., the falling vertical stroke of the T).

Some Chinese character components conforming to this midpoint-start pattern are shown in (25). There are cases of a stroke ending at the midpoint of another stroke, as in (26), but most of these also conform to the midpoint-start pattern, as in (26b).

(25)   丂刀乃丌才彳夂久攵不牙手毛气牛片斤氏勿尹毋弔
(26) a.   屮厶土士彐幺夬
      b.   工夕亙王歹五止日月及爿丑口田由甲

The explanation for these preferences in stroke direction and contact lies in how strokes are written, and as with the visual patterns observed by Changizi, Zhang, Ye, and Shimojo (ibid.), the motoric constraints are universal. Here the most ambitious survey is van Sommers (1984) (see also the summary in van Sommers, 1989), who reports a series of analyses and experiments on the production of simple line drawings. Regard-

ing individual strokes, writers (and sketchers) prefer to pull the writing
instrument rather than to push it, which means that right-handers, who
dominate in the population, draw strokes rightward and/or downward
(yielding ambiguous preferences for counterdiagonal strokes), though
left-handers often draw strokes leftward and/or downward. The con-
ventions of Chinese stroke direction, prescriptively imposed on left-
handers as well, are thus not arbitrary.

The experiments reviewed in van Sommers (1984; 1989) also confirm
the universality of the midpoint-start pattern of the T configuration,
which has also been noted in many other studies (Goodnow and Levine,
1973; Ninio and Lieblich, 1976; Nihei, 1983; Simner, 1981; Smyth, 1989;
Thomassen and Tibosch, 1991). Of course, as Smyth (1989) points out,
stroke coordination also depends on hand-eye coordination, so this is
not a purely motoric process.

The literature generally describes this interaction as one stroke be-
ing anchored on the other; I will call it midpoint anchoring. As Nihei
(1983) recognizes, midpoint anchoring is distinct from what he calls
fluid anchoring, also called threading (Thomassen and Tibosch, 1991)
or chaining (Myers, 2019), whereby a stroke continues from where the
previous left off, without lifting the writing instrument, as in complex
strokes in Chinese. Like midpoint anchoring, chaining seems quite nat-
ural, appearing in the drawing habits even of very young children; the
high frequency of both the T and L configurations in Changizi, Zhang,
Ye, and Shimojo (2006) may thus have some motoric motivation as well.

Another type of natural interaction is what Nihei (1983) calls fixed
anchoring, where two strokes begin at the same point, something that
children find particularly easy to do. Given the rightward and down-
ward stroke directions, in Chinese components the shared starting point
is always at the upper left, as in (27).

(27)   厂 广 匚 冂 几 又 口

The high frequency of the X configuration suggests that stroke cross-
ing should also be relatively simple, but as the above studies report,
young children sometimes draw it as if it were a set of four strokes with
a common starting point (i.e., using fixed anchoring). Its intermediate
difficulty may arise from needing to coordinate two stroke midpoints
rather than relying on a shared starting point, as in fixed anchoring, or
identifying just one midpoint to use as the starting point for the other,
as in midpoint anchoring.

The most difficult stroke interaction is the one Nihei (ibid.) calls bal-
listic, where one stroke ends at another. As with firing a projectile, here
the writer/sketcher must plan the initial action in order to achieve an
end goal, something that young children have particular trouble with.
Its relative rarity in Chinese components, as suggested by (26) above, is

thus expected (and in the next section I will argue that it is even rarer than it seems).

By way of summary, Table 1 lists various types of motoric stroke interactions with associated visual configurations and some Chinese examples.

Table 1. Basic stroke interactions

Interaction	Configuration	Example
None	I	二
Fixed anchoring	L	厂
Chaining	L	し
Midpoint anchoring	T	丁
Crossing	X	乂
Ballistic	T	丄

3.2.2. *Basic Principles of Stroke Group Structure*

If stroke groups have syllable structure, their "nuclei" must be obligatory like those in spoken and signed syllables. If we adopt this hypothesis, then, we must view the smallest logically possible stroke groups, namely isolated (non-contacting) strokes, as consisting solely of a nucleus. This conclusion, consistent with the discussion in earlier sections, also links up with the observation that isolated full (non-dot) strokes tend to share axis with the nearest full stroke, as illustrated in (28): total assimilation in spoken and signed languages seems never to occur syllable-internally, only across syllables (e.g., vowel harmony).

(28)   二丿巜彳彡巛川

By the same reasoning, parallel strokes should represent separate syllables even if they make contact with the same stroke, as in (29). This too is consistent with the above discussion, where we saw that stroke contact of this type does not prevent curving or prominence, both diagnostics for separate stroke groups.

(29)   干土王夫牛兀廾卅井

In spoken syllables, nuclei are obligatory because they represent sonority peaks, making them a plausible candidate for the articulatory target of the entire syllable gesture. In articulatory experiments on American English speech, for instance, Browman and Goldstein (1988) found that the temporal duration remained relatively constant from the midpoint of an onset cluster to the nucleus in the same syllable, regard-

less of the size of the cluster. Speakers thus seem to work with a mental clock that is defined in terms of syllable-internal gestures. Nevertheless, as has often been noted (e.g., Prince and Smolensky, 2004), syllable inventories and prosodic processes both favor onsets and disfavor nucleus-initial syllables. It thus seems reasonable to suppose that when an onset is present (as it is most of the time), the timing of the nucleus depends on it rather than the other way around.

In T configurations, the onset analog would then be the stroke whose midpoint provides the starting point for the other, the analog of the nucleus. Only if the writer intends to write just one stroke is it conceptualized as a nucleus (this conceptual flip is possible because of the lack of intrinsic sonority in strokes). These analyses are sketched in (30), with O for onset and N for nucleus.

(30)   a.   ⼀   ⊤
        b.   N   ON (O = ⼀, N = │)

The proposed contrast can be made more explicit, as in Figure 2, using an autosegmental syllable model that includes moras. Here these structures are interpreted as stating that in T configurations, the location of the nucleus (μ) depends on that of the syllable as a whole (σ), which is assigned by the onset if present.
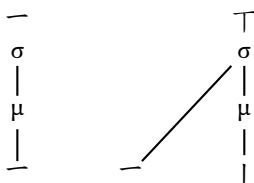
FIGURE 2. Autosegmental analyses of isolated stroke and T configuration

Since character components, as grammatical entities, are amodal, uniting motoric and visual aspects, we should not require that the sequence of strokes or stroke groups in analyses like Figure 2 must correspond with stroke order. Instead the order should be whatever makes the overall analysis the simplest. Thus even though the strokes and stroke order in the components in (31) are identical (left diagonal first), the strokes differ in interaction roles (i.e., which one provides the midpoint anchor). This allows us to express the contrast as in Figure 3, using the same strokes and abstract syllable structures, but different autosegmental links (the contrast is clearer in typefaces that mimic handwriting). As Myers (2019, Section 3.6.2) argues for numerous other reasons, stroke order should be considered part of the articulatory phonetics of character grammar, not part of character phonology per se (e.g., stroke order is surprisingly variable both within and across writers, while character form is much more stable).

(31) a. 人
      b. 入
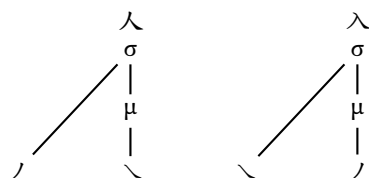
FIGURE 3. Autosegmental analyses of contrasting diagonal T configurations

The autosegmental framework allows us to express other types of stroke interactions as well. In contrast to the T configuration, the X configuration involves two strokes that share a single location. Conceptually, in producing a cross as in (32a), the writer is trying to place two strokes, with distinct axis features, in the same place. This situation may be codified as in (32b), as a single syllable with a short nucleus (N rather than NN), or more explicitly as in Figure 4, with the two strokes linked to a single mora.

(32) a. 十
      b. N
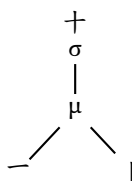
FIGURE 4. Autosegmental analysis of X configuration (crossed strokes)

Since parallel strokes can only appear in separate syllables, in more complex stroke combinations each T and X configuration must form a separate syllable as well, with the syllabic affiliations of the shared stroke(s) indicated through autosegmental association lines. The components in (33a), then, have the syllable structures represented linearly in (33b) (with syllable boundaries marked "."), and autosegmentally in Figure 5. The cross-syllable association lines are dotted to indicate that they do not actually intersect with the others; each syllable is meant to be lying in its own plane.

(33) a. 兀   艹   千   井
      b. ON.ON   N.N   ON.N   N.N.N.N
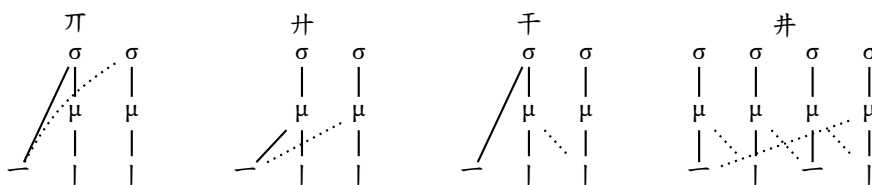
FIGURE 5. Autosegmental analyses of combinations of T and X configurations

One nice consequence of the analysis so far is that by treating the midpoint-anchored stroke in T configurations and crossed strokes in X configurations as nuclei (linked to moras), it puts them in the same class as isolated strokes. As we saw in section 2.3, strokes that are free at their endpoint (i.e., crossed or midpoint-anchored strokes) are subject to prominence and curving, just like isolated strokes. If prominence is an analog of stress and curving an analog of vowel reduction, it makes sense that both would be consistently realizable on the analog of the nucleus.

While midpoint anchoring involves a stroke-on-stroke dependency and crossing involves a symmetrical inter-stroke relationship, strokes sharing a fixed anchor refer to a point that is external to both. It is thus possible to see such strokes as sharing a single empty onset slot (for empty onsets in spoken language, see Marlett and Stemberger, 1983). This would make both strokes themselves into nuclei, as sketched in (34) and Figure 6, with the empty set symbol representing the featureless onset.

(34)   a. $\ulcorner$
        b.   ON.ON
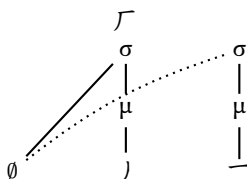
FIGURE 6. Autosegmental analysis of fixed anchoring

Even though the horizontal stroke in (34) starts from the vertical stroke, the latter is not itself an onset, but the nucleus (linked to a mora) in a separate syllable. This is why it may be curved (i.e., undergo a prosodic process akin to vowel reduction). A stroke undergoing left edge curving may also be crossed, as in (35), because crossing strokes are also moraic.

(35)   a.   丗卅
       b.   力九尹

However, if curved strokes are nuclei, they should be incapable of serving as the onset for midpoint anchoring, since onsets are linked directly to the syllable node and have no mora. Yet as the examples in (36) suggest, T configurations do sprout from curved strokes in a small number of components.

(36)   片 月

Perhaps in such rare cases, the leftmost stroke is both the nucleus of one syllable and the onset for another, a situation that can indeed arise in spoken languages (see, e.g., Dell and Elmedlaoui, 1988). This would result in the linear analysis for two of the strokes in (37a) given in (37b), with the autosegmental structure as in Figure 7. Since curving itself is partly lexicalized (see section 2.3), perhaps this unusual syllable structure is as well.

(37)   a.   片 (ノ 一 portion)
       b.   N.ON...
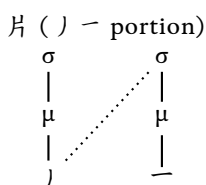
FIGURE 7. Autosegmental analysis of curved stroke offering midpoint anchoring

The next stroke interaction to consider is the chaining of simple strokes to form a complex stroke, as in (38) (these are all of the complex strokes that are also considered to be components by Chuang and Teng, 2009). Since complex strokes are still single strokes, they should be analyzed as comprising a single stroke group. The simplest analysis would thus be to treat all simple segments within a complex stroke as part of the nucleus, that is, as a separate mora.

(38)   a.   乚乛乙乚乀乁乀
       b.   乙乚乙

This analysis is illustrated linearly in (39) and autosegmentally in Figure 8. Note that by giving each stroke segment its own mora, we

capture the observation that complex strokes tend to take up more space than the I, T, and X configurations, all of which are analyzed as monomoraic.

(39) a. ⎤
      b. NNN
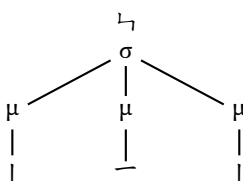
FIGURE 8. Autosegmental analysis of a complex stroke

Although trimoraic syllables like that posited above are rare in spoken languages, they are not impossible, and such complex strokes tend to be disfavored in Chinese character components as well; Chuang and Teng (2009) report lower type frequencies for the components in (38b) as compared with those in (38a).

The last stroke interaction to analyze is the ballistic interaction, where one stroke ends at another. Recall that this interaction is hard for children to learn and relatively rare in character components. If the ease and high frequency of fixed and midpoint anchoring relate to onsets being unmarked in syllables, the markedness of the ballistic interaction suggests that it may relate to the most marked syllable component, the coda, which is as disfavored in languages as the onset is favored (e.g., Prince and Smolensky, 2004). Crucially, the source of the markedness seems similar as well: like ballistic strokes, timing the coda properly requires planning ahead. For example, Browman and Goldstein (1988) found that while American English speakers coordinated the nucleus with the onset cluster as a whole, each of the individual coda consonants were coordinated separately with each other.

Not all end stroke contact is coda-like, however. Since the bottommost stroke in (40a) undergoes prominence, it must be a separate stroke group, and thus cannot form the coda for the other strokes, even though there is a ballistic interaction. A similar conclusion applies in (41), given the larger size of the right-edge stroke. Such contact is thus posited to involve cross-syllabic coordination (like cross-syllabic assimilation in stroke axis), rather than syllable-internal structure. As promised earlier in section 3.2.1, then, what seem to be ballistic interactions in character components are often merely closely concatenated but separate stroke groups.

(40) a. 工
      b. ON.N
(41) a. ㅓ
      b. NN.N

By contrast, as noted in section 3.1, the bottommost stroke in (42a) remains short because it is not free on its endpoint. Here, then, we have a plausible candidate for a coda analog, resulting in the syllable-final structure indicated in (42b), with a long nucleus (the complex stroke) plus coda.

(42) a. 口 (丨 フ 一)
      b. ...NNC (フ 一)

A spot of bother is presented by the first stroke in this component. Even though the left edge stroke shares a fixed anchor with the complex stroke, a situation that we analyzed above as disyllabic sharing of a single empty onset, examples like that in (43) remind us that this entire complex can be subject to prominence, and thus must comprise a syllable as a whole. This forces us to treat the left edge vertical stroke as an onset, resulting in the analysis in (44), or in autosegmental terms in Figure 9. Perhaps this is justified because the left edge stroke is also unusual in another way: its endpoint defines the starting point of another stroke, but since they are not produced in sequence, these two strokes are not chained.

(43) 串
(44) a. 口
      b. ONNC
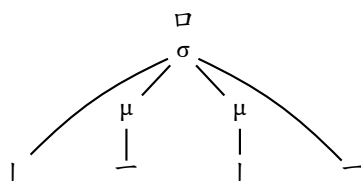
FIGURE 9. Autosegmental analysis of ballistic stroke in a box-shaped stroke group

Note that in Figure 9 I have linked the coda directly to the syllable node. This differs from the more common autosegmental representation for closed syllables (as in Figure 1 above), where the coda is linked to a mora (e.g., Hayes, 1989). Nevertheless, direct linking of the coda to the syllable node has also been argued for in spoken phonology (e.g., Tranel, 1991). While this representation implies that the nucleus and

coda do not form a constituent (the rime), syllables need not have rimes; sign languages have syllables (Sandler, 2008; Sandler and Lillo-Martin, 2006) but I am unaware of any argument that they have rimes, and even some spoken languages provide at best only weak evidence for them (Berg and Koops, 2010).

Moreover, directly linking the coda to the syllable is needed here to avoid ambiguities in interpretation. We have already decided that each of the simple segments in a complex stroke links to its own mora, so giving the coda a mora here would falsely imply a three-segment complex stroke. Alternatively, letting it share the mora with a nuclear stroke would falsely imply crossed strokes. This analytical situation ultimately arises from the lack of intrinsic stroke sonority, which forces the moraic structure itself to do all of the work.

The non-moraic coda hypothesis does have some advantages, however. One is that it helps capture the fact that the ballistic stroke not only ends at another stroke, but also starts at another, namely the leftmost stroke that we analyze as the onset. By linking both the onset and coda to the same node (σ) we imply that they share a location as well. Indeed, as we saw earlier in section 3.2.1, ballistic strokes often start at a leftmost vertical stroke; further examples are given in (45).

(45)   尸 日 目 月

Another advantage is that the non-moraic coda keeps the stroke group small, as with the I, T, and X configurations, in contrast to the the polymoraic representations posited for the larger complex strokes. This point is illustrated by the compact components in (46).

(46)   a.   日   目
        b.   ONNCC   ONNCCC

The analysis also merges naturally with the one given above for curved strokes that act as midpoint anchors. As shown in the autosegmental representation of (47) given in Figure 10, it is straightforward to indicate that this stroke performs double duty as nucleus of one syllable and onset for another.

(47)   a.   月
        b.   N.ONNCC

The autosegmental analysis of stroke sharing introduced in (33) and Figure 5 also allows us to treat the two box-shaped structures in (48) as separate stroke groups, necessary to explain how only the lower one is subject to prominence; see Figure 11.
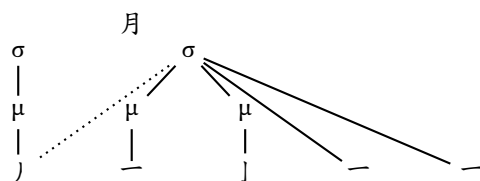
FIGURE 10. Autosegmental analysis of a ballistic stroke starting from a curved stroke

(48) a. 吕 (as in 官)
   b. ONNC.ONNC
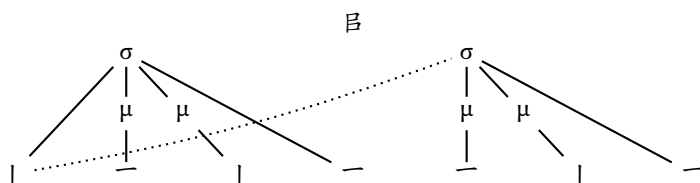
FIGURE 11. Autosegmental analysis of onset stroke shared by two box-shaped stroke groups

The coda analysis also seems appropriate for box-like structures containing a ballistic stroke but missing one or more sides. For example, the top portion of the component in (49a) seems analyzable as indicated by the underlined portion of (49b), where the onset is the curved vertical stroke at the left, the nucleus is the horizontal stroke starting from it (completing the T configuration), and the coda is the short vertical stroke at the upper right that makes endpoint contact.

(49) a. 片
      b. N.ONC.ONN

Ballistic strokes may appear in onsetless syllables as well, however. The cases in (50) can be analyzed in terms of cross-syllable contact between two nuclei rather than a coda (as in (40) and (41) above). Namely, in (50a) and (50b) the contacted stroke is prominent (lengthened) and reduced (curved), respectively, both hallmarks of independent stroke groups.

(50) a. 爿 止
      b. 非

The cases in (51), however, do not show clear signs of the contacted stroke being in a separate syllable. For example, the character in (51c) contains two Chuang and Teng (2009) components, where that on the left (which lacks a Unicode entry) has two ballistic strokes ending at a vertical stroke and that on the right has one ballistic stroke running leftward and downward into the vertical segment of a complex stroke. In none of these cases is there clear prominence or curving in the stroke providing endpoint contact.

(51)   a.   彐臼
       b.   雪 (bottom component)
       c.   北

In such cases, linking the coda strokes to the syllable node does not imply that it starts at the onset, simply because there is no onset, as indicated in (52) and Figure 12.

(52)   a.   北 (left component)
       b.   NCC
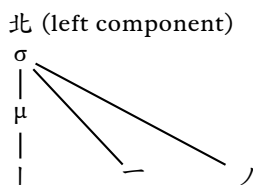
FIGURE 12. Autosegmental analysis of ballistic strokes without starting point contact

It should be clear by now that our neat inventory of basic stroke interactions cannot hope to cover all of the attested interactions that arise as the number and complexity of strokes increases. Changizi, Zhang, Ye, and Shimojo (2006) managed to restrict the scope of their investigation to just 36 visual configurations by imposing a maximum of three strokes, and then restricted it further by considering topology rather than geometry. By contrast, the Chuang and Teng (2009) inventory has 441 geometrically distinct Chinese character components containing up to 17 strokes.

Unsurprisingly, then, from now on the analytical problems and attempted technical fixes come fast and furious, so before continuing I offer the reader a last peaceful moment in the form of Table 2, which summarizes the proposed syllable structures for various types of simple stroke interactions.

Table 2. Basic stroke interactions

Interaction	Example	Syllable structure
None	二	N.N
Fixed anchoring	厂	ON.ON
Chaining	ㄴ	NN
Midpoint anchoring	丁	ON
Crossing	乂	N
Ballistic to prominent/curved stroke	丄	N.N
Ballistic to non-prominent/curved stroke	口	…NC

3.2.3. More Complex Stroke Interactions

Space (fortunately) precludes a complete analysis for each and every character component, and the complex ways in which strokes can interact (unfortunately) precludes a particularly coherent overview. Thus I will merely illustrate a few cases, from what seems to me to be the least to the most problematic.

I start with crossed complex strokes, as in (53). At first it seems it may be hard to specify the precise location of the crossing, but as seen in Figure 13, we can easily code the two complex strokes via bimoraic syllables and the crossing via association lines linking the appropriate simple strokes to a single mora.

(53) a. 乜
      b. NN.NN
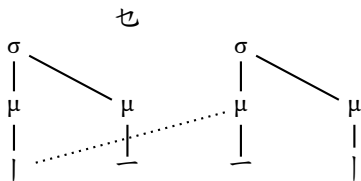
Figure 13. Autosegmental analysis of crossed complex strokes

A slightly trickier situation arises in (54), where a complex stroke not only crosses another stroke, but also shares its starting point. Consistent with the analyses in section 3.2.3, the two strokes must thus also share an empty onset slot, as in Figure 14.

(54) a. 又
      b. ON.ONN
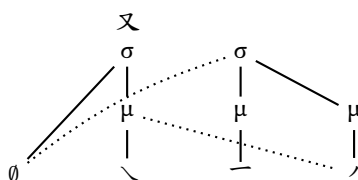
FIGURE 14. Autosegmental analysis of crossing strokes sharing a starting point

In section 3.2.3 we saw that left edge curved strokes, hypothesized to be syllable nuclei, can nevertheless provide the starting point in midpoint anchoring, forcing us to treat them as onsets as well. A similar ambiguity in syllable position arises with complex strokes. Even though each segment in a complex stroke is moraic, it is still possible for a segment to offer midpoint anchoring, as in (55) (the third example contains two Chuang and Teng, 2009, components because the relevant component has no Unicode entry). Autosegmentally this can be handled by doubly linking the segment that serves both as anchor and as part of the complex nucleus, as in Figure 15.

(55)   a.   刀   乃   与
       b.   NN.ON   NNNN.ON   (N.)NNN.ON
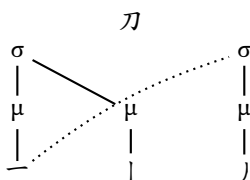
FIGURE 15. Autosegmental analysis of midpoint anchoring from a complex stroke

As we already saw in the previous section, some of the greatest challenges come from the analysis of ballistic strokes, since while by definition they end at another stroke, they typically also start from another stroke, making their position within the stroke group analytically ambiguous. As illustrated in (56), the starting point may even be the first segment of a complex stroke.

(56)   a.   皿 (│ ㄱ │ │ 一)
       b.   罒 (│ ㄱ │ │ 一)

Conveniently, the prominence of the bottom stroke in (56a) shows that it is a separate stroke group. We have also already just seen that

midpoint anchoring from one segment of a complex stroke can be ana-
lyzed as an onset-nucleus structure (here repeated twice, one per inter-
nal stroke). All of these considerations lead to the linear analysis in (57)
and autosegmental representation in Figure 16. Aside from the highly
counterintuitive idea that such a small component could really contain
so many stroke groups, there is no technical problem yet.

(57)   a.   罒

       b.   ON.ONN.ON.ON.N
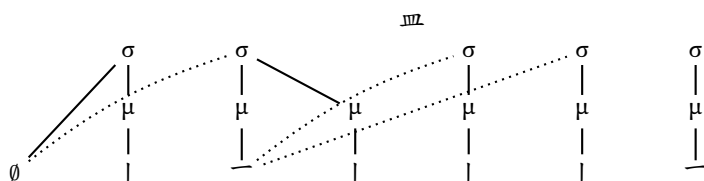
FIGURE 16. Autosegmental analysis of midpoint anchoring from a complex stroke

The component in (56b), however, appears to be a single box-shaped
stroke group. We are thus obliged to somehow represent the anchor-
ing of the two internal strokes from the top right complex stroke while
still recognizing them as ballistically ending at the bottommost stroke
within the same syllable. While it is trivial to give it the same linear
analysis as we did for its rotated counterpart, as in (58), the autoseg-
mental representation in Figure 17 seems to falsely imply that two of
the coda strokes start at the left stroke (since they all directly link to
the syllable node), whereas actually only one of them does (namely the
stroke forming the box bottom). Perhaps we could stipulate that if on-
sets and codas are identical stroke types they cannot be interpreted as
linked together; stroke contact requires a difference in axis. Even so, this
analysis has the additional counterintuitive effect of giving this compo-
nent a totally different structure from the virtually identical component
in (57).

(58)   a.   皿   目

       b.   ONNCCC   ONNCCC

A particularly striking example of the challenges posed by my analy-
sis of ballistic strokes as coda-like is the component in (59a), which con-
tains two horizontal ballistic strokes linking two vertical ones (plus a
fifth forming the bottom of the box). The T configuration at the top is
readily analyzed as ON, but simply concatenating all five ballistic strokes
as codas, as in (59b), fails to indicate which stroke links with which.
Nevertheless, given that the lower box seems prominent as a whole, the
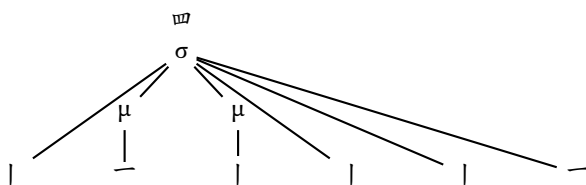
FIGURE 17. Autosegmental analysis of ballistic stroke in a box-shaped stroke group

treatment here of it as a single stroke group does at least capture that prosodic observation. Moreover, the oddity of this type of situation is correlated with its rarity in the Chuang and Teng (2009) inventory.

(59)   a.   面
       b.   ON.ONNCCCCC

The component family in (60) raises further problems (those in (60e-f) are not in the Chuang and Teng, 2009, component inventory but are included for completeness).

(60)   a.   田
       b.   由
       c.   甲
       d.   申
       e.   甶
       f.   甲

The box seems to form a single stroke group because in (60b-c) it is the target of bottommost prominence, with the extended stroke as a separate stroke group. That makes the internal horizontal stroke part of the same stroke group as the box, and thus a coda, as in earlier analyses. The problem is that this stroke is also crossed, which means it is moraic, but codas cannot be moraic (or else the autosegmental representations become ambiguous, as discussed earlier).

One way to respond to this challenge is to start with the supposition that the box-shaped stroke group is actually that in (61a), as analyzed in (61b), whereas the central vertical stroke is a separate stroke group in all cases, including in (60a). The crossing problem can then be dealt with by treating crossing here as a mere accident of the central vertical stroke's starting and ending points, rather than being something represented phonologically. This seems counterintuitive given the high salience of the cross, but if the coiners of (60e-f) were able to decompose it, doing so is not impossible.

(61) a. 曰
      b. ONNCC

This now merely leaves us with the challenge of representing the vertical stroke's position within the formal straightjacket I have set myself. Across the components in (60), the starting point of this stroke is variously above the box, at the top of the box, or at the central horizontal stroke, which can be represented respectively as an onsetless syllable, as a syllable with the onset in the first segment of a complex stroke (as in (55) and Figure 15 above), and as a syllable with the onset at the box's first coda stroke. The ending point of the vertical stroke is variously at the central horizontal stroke, at the bottom of the box, or below the box, which can be represented respectively as sharing a coda with the box's first coda stroke, as sharing a coda with the box's final coda stroke, or as being a codaless open syllable. None of these possibilities raises any fatal problems, as sketched in (62) and (63), with subscripts to indicate the cross-syllable autosegmental linking. Figure 18 spells out the idea for one component.

(62) a. 田   甲   甲
      b. $ON_1NCC_2.O_1NC_2$ $ON_1NCC.O_1N$ $ONNC_1C.O_1N$
(63) a. 甲   由   占
      b. $ONNCC.N$ $ONNCC_1.NC_1$ $ONNC_1C.NC_1$

田
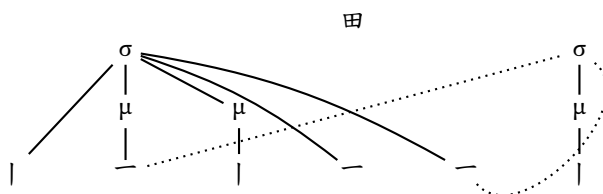
FIGURE 18. Autosegmental analysis of ballistic stroke in a box-shaped stroke group

Another challenging component family is that in (64). In the first component, the lower complex stroke shares its start with the ballistic stroke, which can be expressed via a shared autosegmental link between the first stroke group's coda and the second stroke group's onset, as indicated by the coindexing; note that neither onset nor coda is moraic, so there is no risk of misinterpreting the shared link as stroke crossing. In the second component, a single onset is shared between the upper and lower complex strokes. In the third component, neither complex stroke has an onset; the ballistic stroke does, but the representational scheme does not allow me to represent it unambiguously so I leave it out, as I

did with the crossed strokes in the previous component family. While hardly an ideal solution, at least all of these representations show the lower complex stroke as forming a separate stroke group, allowing it to be subject to bottommost prominence.

(64) a. 己   巳   已
     b. $NNC_1.O_1NN$ $O_1NNC.O_1NN$ $NNC.NN$

I end my survey with an analysis of the most complex component in the Chuang and Teng (2009) inventory, that in (65a). Figure 19 indicates schematically which component parts correspond to which of the stroke groups listed in (65b).

(65) a. 龜
     b. ONN.NNC.NNC.NNC.ONNCC.ONC.N.ONN

FIGURE 19. Sketch of a stroke group analysis of the most complex component

While no utterly fatal problems have arisen in this survey, we have needed a plethora of special devices (if not special pleading), some of which have yielded counterintuitive results. More importantly, I have yet to provide any argument that any of this matters to actual readers or writers. Collecting proper psycholinguistic data will have to await the proverbial future research, but in the next section I do examine one possible psychological implication of the stroke-group-as-syllable analysis.

3.3. Stroke Groups and Menzerath's Law

A demonstration that stroke groups, as I have identified them, conform to the Menzerath-Altmann law would be consistent with the claim that they influenced the evolution of characters into their modern forms. The modeling work here is based on syllabic analyses for all 441 components of Chuang and Teng (ibid.)[2]. The analyses are based on component forms as they appear in Chuang and Teng's regular (handwriting

---

     2. The data are available at https://osf.io/nbhcm/.

style) typeface. A variety of analytical decisions are scattered through-
out, and I am not entirely sure if I have applied all of my principles
completely consistently, but hopefully this merely added noise and not
bias.

If stroke groups have some validity, we expect that within charac-
ter components, there should be an inverse power relationship between
mean stroke group complexity and the number of stroke groups. To test
this, I operationalized stroke group complexity as the number of O, N,
C segments in the linear syllabic analyses, where N represents a mora
and O and C represent simple strokes without a mora. Autosegmental
lines are not counted.

Following Prün (1994), Figure 20 shows the nonlinear best-fit for the
simplified Menzarath equation in (66a), with the model parameters and
other statistical values shown in (66b).

(66)   a. $y = ax^b, b < 0$
        b. $a = 2.44, b = -0.19, p_b < .0001, R^2 = .82$

The coefficients are of the expected signs (positive a, negative b) and
statistically significant; here I highlight p_b, the p value for b, which con-
firms that this is an inverse power function (against the null hypothesis
$b = 0$). However, the data points are much more scattered than in other
applications of Menzerath's law to Chinese script. Again following Prün
(ibid., p. 149), I quantified model fit using the coefficient of determina-
tion R^2 (Prün labels it D). As shown in (66b), this value is relatively high
but still far below the $R^2 = .99$ reported by Prün (ibid.) for component
complexity in characters.
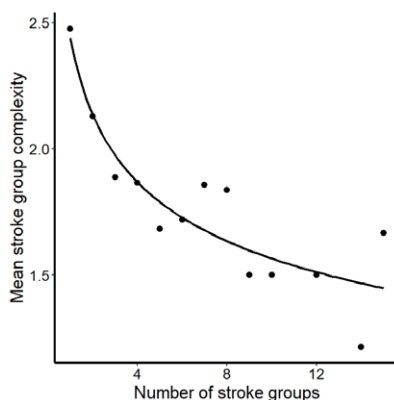
FIGURE 20. Mean stroke group complexity as a function of the number of stroke
groups

While the less than perfect fit may relate to inconsistencies in how stroke groups were identified, could the fact that there is any fit at all be dismissed as confounding with other factors known to obey Menzerath's law? In particular, Bohn (1998) demonstrated an inverse power relationship between mean stroke complexity and stroke number within character components. In my analyses, isolated and crossed simple strokes are the simplest possible stroke groups (N), whereas complex strokes are necessarily more complex (NN...). Thus it could be that Figure 20 merely recapitulates Bohn's analysis in an obscured form.

The ideal way to rule this out would be to build a model that also includes stroke number and stroke complexity as interacting factors, but statistical interaction is only well defined for (generalized) linear models, not the nonlinear model that I used to fit the power law. However, it is still possible (as well as conceptually simpler and less assumption-prone) to build separate nonlinear models for multiple subsets of the data, in each of which stroke number and complexity are held constant. If Menzerath's law still applies in each of the subsets, this cannot be ascribed to stroke number or complexity.

This I did for eight subsets of components with two to five strokes, where the mean stroke complexity (as defined by Bohn, 1998, where hooks add complexity) was either 1 or higher than 1 (ranging from 1.2 to 3); outside these ranges the subsets were too small, falling below ten data points per subset. These subsets still cover the majority of the total data (over 65%). As can be seen in Table 3, all of the subsets are consistent with an inverse power law, though not all are statistically significant or show very strong model fits.

TABLE 3. Menzerath's law across subsets of character components

	Mean stroke complexity $= 1$ Stroke number				Mean stroke complexity > 1 Stroke number			
	2	3	4	5	2	3	4	5
a	1.64	2.84	3.88	2.51	3.00	3.76	5.06	6.06
b	−0.57	−0.97	−1.02	−0.44	−0.56	−0.84	−0.84	−0.86
p_b	0.013	<0.0001	<0.0001	0.23	0.0001	<0.0001	<0.0001	<0.0001
R^2	0.30	0.49	0.59	0.14	0.27	0.41	0.72	0.76

The results thus add some (weak) support for the claim that stroke groups may be psychologically real, or at least were while characters were evolving.

4.   Conclusions

There is no doubt that Chinese character components and strokes are psychologically real levels of structure. It is also relatively easy to see

components as analogous to morphemes and strokes as analogous to phonological segments. The evidence for an intermediate syllable-like level, the stroke group, is nowhere near as strong, but this initial exploration has nevertheless uncovered some interesting patterns. Compared with the list of syllable properties in (19) above, the stroke group scorecard in (67) suggests that there may indeed be some genuine similarities with syllables, as marked in italics.

(67)   a.   In a stroke group, there is no analog to intrinsic sonority.
       b.   There is as yet no evidence that stroke groups are perceptually salient.
       c.   *Stroke groups are targeted by analogs to foot-level processes like stress.*
       d.   *Articulatory gestures are more closely coordinated within than across stroke groups.*
       e.   *Stroke groups may have analogs to obligatory nuclei, favored onsets, and disfavored codas.*
       f.   *Stroke groups compete for space in character components (Menzerath's law).*

Future research can take many possible directions. The most fundamental question is whether the syllable analogy is really needed to preserve whatever is genuine in the stroke group hypothesis. After all, even in sign language phonology there have been arguments that what most linguists consider to be syllables may actually be more analogous to complex segments (Channon, 2003). Another issue is how to extend the present analysis to the many writing systems that, unlike Chinese script, have strongly curved strokes, including circles, semicircles, and loops; among these are systems historically derived from Chinese, like Japanese hiragana. Previous analyses of Roman letters have considered curved strokes (e.g., Watt, 1980; Primus, 2004), and van Sommers (1984) includes a chapter-length discussion of the production of curvilinear forms. Still, circle-like strokes do complicate matters, particularly since they allow two strokes to contact each other in more than one place, a possibility we did not have to worry about when analyzing Chinese script. In my opinion, however, most urgently needed is the collection of psycholinguistic evidence that writers and readers actually do learn or process characters in terms of stroke groups.

Regardless of how research progresses, I hope this preliminary study has at least revealed the rich and challenging nature of a still underexplored aspect of writing systems: the precise formulation of stroke interactions.

References

Altmann, Gabriel (1980). "Prolegomena to Menzerath's law." In: *Glottometrika* 2.2, pp. 1–10.

Aronoff, Mark (1994). *Morphology by itself: Stems and inflectional classes*. Cambridge, MA: MIT Press.

Beckman, Mary E and Jan Edwards (1990). "Lengthenings and shortenings and the nature of prosodic constituency." In: vol. 1. Papers in laboratory phonology. Cambridge, UK: Cambridge University Press, pp. 152–178.

Behr, Wolfgang (2006). "'Homosomatic juxtaposition' and the problem of 'syssemantic' (*huìyì*) characters." In: *Écriture chinoise: données, usages et représentations*. Ed. by Françoise Bottéro and Redouane Djamouri. Paris: École des hautes études en sciences sociales, Centre de recherches linguistiques sur l'Asie orientale, pp. 75–114.

Berent, Iris and Amanda Dupuis (2017). "The unbounded productivity of (sign) language: Evidence from the Stroop task." In: *The Mental Lexicon* 12.3, pp. 309–341.

Berg, Thomas and Christian Koops (2010). "The interplay of left-and right-branching effects: A phonotactic analysis of Korean syllable structure." In: *Lingua* 120.1, pp. 35–49.

Bohn, Hartmut (1998). *Quantitative Untersuchungen der modernen chinesischen Sprache und Schrift*. Hamburg: Kovač.

Browman, Catherine P. and Louis Goldstein (1988). "Some notes on syllable structure in articulatory phonology." In: *Phonetica* 45.2-4, pp. 140–155.

Changizi, Mark A. et al. (2006). "The structures of letters and symbols throughout human history are selected to match those found in objects in natural scenes." In: *The American Naturalist* 167.5, E117–E139.

Channon, Rachel Elizabeth (2003). "Signs are single segments: Phonological representations and temporal sequencing in ASL and other sign languages." PhD thesis. College Park: University of Maryland.

Chen, Heng and Haitao Liu (2019). "A quantitative probe into the hierarchical structure of written Chinese." In: *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pp. 25–32.

Chen, Jenn-Yeu and Rong-Ju Cherng (2013). "The proximate unit in Chinese handwritten character production." In: *Frontiers in Psychology* 4. DOI: 10.3389/fpsyg.2013.00517.

Chuang, D.M. [莊德明] and H.Y. Teng [鄧賢瑛] (2009). 漢字構形資料庫的研發與應用 *[Research and development of Chinese characters information database and its application]*. 臺北市 [Taipei]: 中央研究院 [Academia Sinica].

Crasborn, Onno A, Els van der Kooij, and Johan Ros (2012). "On the weight of phrase-final prosodic words in a sign language." In: *Sign Language & Linguistics* 15.1, pp. 11–38.

Cutler, Anne and Charles Clifton (1984). "The use of prosodic information in word recognition." In: *Attention and performance X: Control of language processes*. Ed. by H. Bouma and D. G. Bouwhuis. Hillsdale, NJ: Lawrence Erlbaum, pp. 183–196.

Dell, François and Mohamed Elmedlaoui (1988). "Syllabic consonants in Berber: Some new evidence." In: *Journal of African Languages and Linguistics* 10.1, pp. 1–17.

Evertz, Martin (2018). *Visual prosody: The graphematic foot in English and German*. Berlin: Walter de Gruyter.

Feldman, Laurie Beth and Witina WT Siok (1999). "Semantic radicals contribute to the visual identification of Chinese characters." In: *Journal of Memory and Language* 40.4, pp. 559–576.

Fujiwara, Yoshi, Yasuhiro Suzuki, and Tomohiko Morioka (2004). "Network of words." In: *Artificial Life and Robotics* 7.4, pp. 160–163.

Goodnow, Jacqueline J and Rochelle A Levine (1973). " 'The grammar of action': Sequence and syntax in children's copying." In: *Cognitive Psychology* 4.1, pp. 82–98.

Haralambous, Yannis (2013). "New perspectives in Sinographic language processing through the use of character structure." In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Vol. 7816. Lecture Notes in Computer Science. Heidelberg: Springer, pp. 201–217.

Hayes, Bruce (1989). "Compensatory lengthening in moraic phonology." In: *Linguistic Inquiry* 20.2, pp. 253–306.

Hue, Chih-Wei (2003). "Number of characters a college student knows." In: *Journal of Chinese Linguistics* 31.2, pp. 300–339.

Ladd, D Robert (2014). *Simultaneous structure in phonology*. Oxford: Oxford University Press.

Levelt, Willem JM, Ardi Roelofs, and Antje S Meyer (1999). "A theory of lexical access in speech production." In: *Behavioral and Brain Sciences* 22, pp. 1–38.

Li, Jianyu and Jie Zhou (2007). "Chinese character structure analysis based on complex networks." In: *Physica A: Statistical Mechanics and its Applications* 380, pp. 629–638.

Lu, Qin et al. (2002). "Decomposition for ISO/IEC 10646 ideographic characters." In: *Proceedings of Conference on Computational Linguistics 2002*.

Marlett, Stephen A and Joseph Paul Stemberger (1983). "Empty consonants in Seri." In: *Linguistic Inquiry* 14.4, pp. 617–639.

McCarthy, John J. and Alan Prince (1998). "Prosodic morphology." In: *The handbook of morphology*. Ed. by A. Spencer and A. M. Zwicky. Oxford: Blackwell, pp. 283–305.

Menzerath, Paul (1954). *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümmler.

Myers, James (2016). "Knowing Chinese character grammar." In: *Cognition* 147, pp. 127–132.

Myers, James (2019). *The grammar of Chinese characters: Productive knowledge of formal patterns in an orthographic system*. London: Routledge.

—————— (2020). "Grapheme size is processed like stress: Experimental evidence from Chinese script." In: *Proceedings of the 53rd Annual Meeting of the Societas Linguistica Europaea (SLE 2020)*. URL: https://osf.io/y6ajz/ (visited on 2020/09/06).

Myers, James and Jane Tsay (2015). "Trochaic feet in spontaneous spoken Southern Min." In: *Proceedings of the 27th North American Conference on Chinese Linguistics*. Vol. 27. 2. Los Angeles: UCLA, pp. 368–387.

Nihei, Yoshiaki (1983). "Developmental change in covert principles for the organization of strokes in drawing and handwriting." In: *Acta Psychologica* 54, pp. 221–232.

Ninio, Anat and Amia Lieblich (1976). "The grammar of action: 'Phrase structure' in children's copying." In: *Child Development* 47.3, pp. 846–850.

Peng, Xuanwei (2017). "Stroke systems in Chinese characters: A systemic functional perspective on simplified regular script." In: *Semiotica* 2017.218, pp. 1–19.

Primus, Beatrice (2004). "A featural analysis of the Modern Roman Alphabet." In: *Written Language & Literacy* 7.2, pp. 235–274.

Prince, Alan and Paul Smolensky (2004). *Optimality theory: Constraint interaction in generative grammar*. Malden MA: Blackwell.

Prün, Claudia (1994). "Validity of Menzerath-Altmann's law: Graphic representation of language, information processing systems and synergetic linguistics." In: *Journal of Quantitative Linguistics* 1.2, pp. 148–155.

Sandler, Wendy (1993). "A sonority cycle in American Sign Language." In: *Phonology* 10.2, pp. 243–279.

—————— (2008). "The syllable in sign language: Considering the other natural language modality." In: *The syllable in speech production*. Ed. by Barbara L. Davis and Krisztina Zajdó. New York: Lawrence Erlbaum, pp. 379–408.

Sandler, Wendy and Diane Lillo-Martin (2006). *Sign language and linguistic universals*. Cambridge UK: Cambridge University Press.

Simner, Marvin L. (1981). "The grammar of action and children's printing." In: *Developmental Psychology* 17.6, pp. 866–871.

Slaměníková, Tereza (2018). "On the nature of unmotivated components in modern Chinese characters." In: *Proceedings of Graphemics in the 21st century*. Ed. by Yannis Haralambous. Vol. 1. Grapholinguistics and its applications. Brest: Fluxus Editions, pp. 209–226.

Smyth, Mary M. (1989). "Visual control of movement patterns and the grammar of action." In: *Acta Psychologica* 70.3, pp. 253–265.

Taft, Marcus and Xiaoping Zhu (1997). "Submorphemic processing in reading Chinese." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23.3, pp. 761–775.

Thomassen, Arnold J.W.M. and Hein J.C.M. Tibosch (1991). "A quantitative model of graphic production." In: *Tutorials in motor neuroscience*. Ed. by G. E. Requin J. Stelmach. Dordrecht: Springer, pp. 269–281.

Tranel, Bernard (1991). "CVC light syllables, geminates and moraic theory." In: *Phonology* 8.2, pp. 291–302.

van Sommers, Peter (1984). *Drawing and cognition: Descriptive and experimental studies of graphic production processes*. Cambridge UK: Cambridge University Press.

———— (1989). "A system for drawing and drawing-related neuropsychology." In: *Cognitive Neuropsychology* 6.2, pp. 117–164.

Wang, Jason Chia-Sheng (1983). "Toward a generative grammar of Chinese character structure and stroke order." PhD thesis. University of Wisconsin, Madison WI.

Wang, Ruiming et al. (2020). "Chinese character handwriting: A large-scale behavioral study and a database." In: *Behavior Research Methods* 52, pp. 82–96.

Wann, John and Ian Nimmo-Smith (1991). "The control of pen pressure in handwriting: A subtle point." In: *Human Movement Science* 10.2-3, pp. 223–246.

Watt, William C (1980). "What is the proper characterization of the alphabet? II: Composition." In: *Ars Semeiotica* 3.1, pp. 3–46.

Viewpoints on the Structural Description of Chinese Characters

Tomohiko Morioka

Abstract. This paper is about our viewpoints and methodology concerning the description of the structure of Chinese characters. First, we describe how components can be detected in characters. When a character is used as a component of a compound character and its shape appears without significant change, then the component can be easily identified. However, in many cases, it is not so easy to find the components that build a character out of purely visual features. One of the factors is simplification of the graphic form when characters are assembled out of components. Such a change of glyph form reduces the connection with pronunciation and meaning of the original character and increases the symbolic aspect of the character. It is particularly complicated when multiple components are combined, transformed and demotivated into a symbolic component. Here we discuss these issues with respect to the productivity of components and to the relationships between components and characters.

1.   Introduction

Most Chinese characters (漢字) can be represented by a combination of components. For example, the character "林" (forest) has the same component "木" (tree) on the left and right, and the character "雲" (cloud) has component "云" (phonetic part) placed under the component "雨" (rain). The structure of Chinese characters, which consists of a combination of components, is not only an abstract expression of its shape, but is also related to its semantic and phonetic values. Therefore, to understand a Chinese character, it is important to find out what components are used, where they are placed in the characters, and how they are combined.

In this paper, such a description of the structure of Chinese character is called a *Hanzi structure description* (漢字構造記述; structure description of Chinese character) (Morioka, 2018b). Various formalisms have been used to describe Hanzi structure. Nowadays a quite widespread formalism is the one of Ideographic Description Sequences (IDS) defined in

Tomohiko Morioka   ⬤ 0000-0001-5315-3383
47 Kitashirakawa-Higashioguramachi, Sakyo, Kyoto, 606–8265, Japan
E-mail: tomo@zinbun.kyoto-u.ac.jp

ISO/IEC 10646 (*Information technology—Universal Coded Character Set (UCS)*) 2014). In this paper, we use the IDS formalism to describe Hanzi structure. For example,

– 林 = ⿰木木
– 雲 = ⿱雨云

in IDS.

In these two examples we can easily detect components and infer structure. However, in some cases, the situation is ambiguous, e.g.,

– 旗 = ⿰方⿱𠂉其 or ⿰⿸𠂆其?
– 嬴 = ⿳亠⿲月女凡 or ⿱羸女?

How do we detect components and the corresponding structure in such cases?

2.   Etymological View

When a character is used as a component of a compound character and its shape is simply inserted into it without significant change, the component can be easily detected via its shape. For example, the character "林" (forest) appears to have two components "木" (tree) arranged side-by-side. Therefore, the structure of "林" can be written as "⿰木木".

Some structures have been preserved even though the graphic forms of Chinese characters have changed significantly over the centuries. For example, "⚎" is an Oracle-bone character corresponding to the modern character "林". Its structure can be written as "⿰❅❅," which is similar to "⿰木木" ("❅" is an Oracle-bone character corresponding to the modern character "木").

Similarly, the character "雲" (cloud) seems to consist of a component "雨" (rain) placed above a component "云" (the structure being "⿱雨云"). The Oracle-bone character "ʒ̄" corresponds to the modern character "雲" *as a character*, but corresponds to "云" as a component of Hanzi structure.[1] That is, it is considered that "雲" was formed by adding the semantic component "雨" to distinguish it from morphemes other than "雲" because "云" was used phonetically. Anyway, in this case as well, the components can be easily found just by looking at the characters.

However, in many cases, it is not so easy to find the components of a character just by examining the graphic form of the character. One of the causes of the problem is simplification of the graphic form of components when they are assembled into characters. For example, "隆" (eminent, exalt) looks like a combination of "阝" and "夆" on the left and right,

---

1. In Mainland China, "云" is a simplified Chinese character corresponding to "雲".

namely "田阝夆". However, "䧙" (Small Seal form (小篆) of Shuowen corresponding to "隆") seems to contain the component "屮" ("生" (be born, living, raw)) inside the component "𨾴" ("降" (fall)), so that the structure could be "田𨾴屮". Compared to the Small Seal form, Hanzi structure of "隆" should be described as "田阝夆生". Through this structural analysis, we see that "阝夆" can be considered as being a simplified form of "降" or a component of it. As described above, there are two approaches to structural description: a structural description based on appearance and a structural description based on etymological explanation.

To adopt the latter position, the etymological knowledge of the Chinese character is required. However, characters with clear etymology are only a small part of the whole, and the etymological data on many characters are unclear or unknown.

3. Component Models

Historically, Hanzi structure descriptions (or the underlying analysis) were written for humans. For example, *Shuowen Jiezi* (説文解字; Shuowen), which is considered to be the oldest radical-based Chinese character dictionary, describes the kinds of components that comprise each compound character. The analysis of Hanzi structures in Shuowen is based on the so-called six-categories classification model (六書) and focusing on components motivated by pronunciation and meaning. In this model, each component is considered to be derived from a character, and each component is considered to (partially) inherit the phonetic and/or semantic value of the original character.

In the twentieth century, Tang Lan (唐蘭) proposed a new research approach and the three-categories classification model (三書). He also focused on graphemes that were not motivated by pronunciation or meaning and named them "symbol characters" (記号字; unmotivated characters), see also Slaměníková (2019). Qiu Xigui (裘錫圭) also made great contributions to the study of symbol characters. Tatsuro Asahara (浅原達郎) greatly contributed to Qiu Xigui's approach by avoiding the classification of components, and by proposing a symbolization (demotivation) model based on a more relational viewpoint (Asahara, 1996). According to this theory, instead of considering the classification of semantic, pronunciation and symbolic components, it is assumed that associative keys connect characters and components motivated by meaning and pronunciation (or act as symbolization filters to remove them).

4. Productivity of Components

Based on Tatsuro Asahara's view, when studying the functionality of a component, it is important to focus on the set of characters that have the same component, and to explore this component's features (or the actual state of componentization) as a common tendency among them. In other words, this can be viewed as a position that focuses on the *productivity of the components*. Tatsuro Asahara gave up explaining the problem of hierarchical componentization, however, by focusing on the productivity of components, it may be possible to identify a composite component acting as a functional unit by analyzing the corpus of the Hanzi structure described visually.

For example, "嬴" can be described as "⿳亡口月女几" (etymological decomposition: "贏" is a phonetic component and "女" (woman) is a semantic component) or as: "⿳亡口⿲月女几" (visual decomposition). When searching for characters containing "贏" in the CHISE-IDS Database[2] (Morioka, 2015), only 50 characters were found in those included in UCS (even limited to components, the following are found: 「嬴」「贏」(「贏」) 「嬴」「嬴」「嬴」「嬴」「嬴」「嬴」「嬴」「嬴」「嬴」「嬴」「嬴」「嬴」「嬴」 「嬴」「嬴」「嬴」). In contrast, only fifteen characters containing "⿲月女几" were found, all of which contained "贏" or some variant of it. Similarly, in the case of "族," many characters containing "𭤴" were found (such as "㐧," "施," "斿," "斿," "旅," "㫃," "㫃," "旍," "㫃," "旗," "旅," "旐," "施," "旋," "旃," "旆," "旒," "旇," "旋," "㫃," "旌," "斿," "虎," "族," "㫃," "㫃," "㫃," "�," ...), whereas only three characters were found that contain "矢".

More detailed results are shown in Tables 1 to 12.

To perform this investigation, we introduced 12 rewriting rules, transforming functional structures into apparent structures. Each table corresponds to one of these rewriting rules:

Rule-111 ⿰⿱ LRB ↔ ⿱ L ⿰ RB
Rule-112 ⿲⿱⿱亻∣ RB ↔ ⿲⿳亻∣⿱ RB
Rule-121 ⿱⿱ ATR ↔ ⿱ A ⿱ TR if T is *tare*
Rule-122 ⿱⿱ ADR ↔ ⿱ A ⿳ DR if D is not *tare*
Rule-131 ⿲⿱⿱ E 亻 R ↔ ⿱ E ⿲亻 R
Rule-132 ⿲⿱⿱ EAB ↔ ⿱ E ⿱ AB if A = non 亻
Rule-210 ⿰⿱⿱ LRB ↔ ⿰⿱⿱ LBR
Rule-411 ⿱⿱ AEM ↔ ⿱ A ⿰ EM if E is *kamae*
Rule-414 ⿱⿱ ABM ↔ ⿱ A ⿱ MB if both A and B are not enclosure
Rule-511 ⿲⿱⿲⿱ LRBA ↔ ⿱⿲⿱ LARB
Rule-611 ⿲⿱ A ⿲ LRC ↔ ⿱ A ⿳ LCR
Rule-612 ⿲⿱ AM ⿲ LRC ↔ ⿱ AM ⿳ LCR

Each table consists of five columns. The first column ("char") display the characters, the second column ("structure") its structure, the third column ("component") the components, and the fourth column (p*n*) indicates the number of character objects in the CHISE character ontology that contain the given *component*. Columns other than *char* are divided into upper and lower subrows. The upper subrow contains information about functional structure and the lower one contains information about apparent structure.

The accuracy value is calculated by the following formula:

$$A_x = \frac{N_{px}^2}{N_p(N_{pf} + N_{pa})} \cdot 100.$$

Let S_{pf} be a set of character objects in the CHISE character ontology that have a functional component, and let $N_{pf} = n(S_{pf})$ (p*n* value of the upper subline).

Similarly, let S_{pa} be a set of character objects in the CHISE character ontology that have a apparent component, and let $N_{pa} = n(S_{pa})$ (p*n* value of the lower subline).

Let $S_p = S_{pf} \cup S_{pa}$, and let $N_p = n(S_p)$.

Let x of A_x and N_{px} be a variable to select *f* (functional) or *a* (apparent).

Note that the CHISE character ontology is based on the Multiple Granularity Hanzi Structure Model (Morioka, 2015; 2018a), so that each number f-p*n* and a-p*n* denotes the plural glyph granularity of Chinese characters such as abstract character, unified-glyph, abstract-glyph (字體), etc. In addition, the CHISE character ontology also includes character objects that cannot be unified by the existing CJKV Unified Ideographs of UCS. However, these tables show only representative glyphs and abstract characters of UCS: <character> indicates abstract characters and characters without angle brackets denote representative glyphs. Some abstract components unify multiple abstract characters, which are expressed as: <人/入/𠆢>.

TABLE 1. ⿲⿸ LRB ↔ ⿰ L ⿱ RB (111)

char	structure	component	p*n*	accuracy
㣇	⿰〈广〉多	〈广〉	167	98.8
	⿰方⿱〈宀〉多	⿱〈宀〉多	1	0.1
旋	⿰〈广〉疋	〈广〉	167	98.8
	⿰方疋	疋	1	0.1

Same as above	牏, 㫍, 㫋, **㫃**, **旗**, 䕐, 㶐, 㡠, 㡾, 旎, 䜣, 㫌, 庵, 㫒, 旀, ⟨䕐⟩, ⟨㫋⟩, ⟨㡠⟩, ⟨㫌⟩, ⟨㡶⟩, ⟨㫊⟩, ⟨䢀⟩, ⟨㫑⟩, ⟨㡾⟩, ⟨㫔⟩, ⟨㫐⟩, ⟨㫒⟩, ⟨㡠⟩, ⟨斒⟩, ⟨㿸⟩, ⟨㡶⟩, ⟨旀⟩, ⟨㡺⟩, ⟨焌⟩, ⟨㫌⟩, ⟨㫂⟩, ⟨㫄⟩, ⟨㫃⟩, ⟨旂⟩, ⟨旆⟩, ⟨㫍⟩, ⟨㫃⟩, ⟨㫐⟩, ⟨㫒⟩, ⟨㡢⟩, ⟨㿶⟩, ⟨㫄⟩, ⟨㫓⟩, ⟨㫑⟩, ⟨㿗⟩, ⟨㫋⟩, ⟨㫌⟩, ⟨㫊⟩, ⟨族⟩, ⟨㡸⟩, ⟨旃⟩, ⟨旄⟩, ⟨旅⟩, ⟨㫂⟩, ⟨㫉⟩, ⟨㫄⟩, ⟨㫃⟩, ⟨旋⟩, ⟨㫀⟩, ⟨㫐⟩, **旛**, ⟨㡾⟩, 㡶, ⟨旂⟩, 㫄, 虎, **旌**, **㫈**, ⟨㡜⟩, ⟨㫒⟩, ⟨旀⟩			
牏	□⟨疒⟩**酉** □**方**⟨酉⟩	⟨疒⟩ ⟨酉⟩	167 2	97.6 0.1
⟨㡝⟩	□⟨疒⟩**言** □**方**□⟨一⟩言	⟨疒⟩ □⟨一⟩言	167 2	97.6 0.1
Same as above	㫄, 㡶, 㫈, 㫄, 㫉, **㫈**, 㫅, 㡿, 㶐, 旈, ⟨旇⟩, ⟨㫃⟩, ⟨㡜⟩, ⟨㫃⟩, ⟨㶐⟩, ⟨㫌⟩, ⟨㫈⟩, ⟨旈⟩, ⟨㫉⟩, ⟨㫓⟩, ⟨旂⟩, **旂**, ⟨㫄⟩, ⟨㫈⟩			
㫋	□⟨疒⟩**冉** □**方**□⟨一⟩**冉**	⟨疒⟩ □⟨一⟩**冉**	167 3	96.5 0.1
⟨旋⟩	□⟨疒⟩⟨正⟩ □**方**𤴓	⟨疒⟩ 𤴓	167 3	96.5 0.1
Same as above	旈, **㫅**, ⟨旇⟩, ⟨旊⟩, ⟨旈⟩			
旋	□⟨疒⟩**正** □**方**𤴓	⟨疒⟩ 𤴓	167 4	95.4 0.1
⟨旗⟩	□⟨疒⟩⟨其⟩ □**方**箕	⟨疒⟩ 箕	167 4	95.4 0.1
族	□⟨疒⟩**矢** □**方**矢	⟨疒⟩ 矢	167 4	95.4 0.1
Same as above	旎, ⟨旇⟩, ⟨㫃⟩, 旂			
㫄	□⟨疒⟩**几** □**方**朰	⟨疒⟩ 朰	167 7	92.1 0.2
⟨旎⟩	□⟨疒⟩**几** □**方**朰	⟨疒⟩ 朰	167 7	92.1 0.2
㫄	□⟨疒⟩**日** □**方**⟨㠯⟩	⟨疒⟩ ⟨㠯⟩	167 10	89.0 0.4
⟨㫄⟩	□⟨疒⟩**日** □**方**⟨㠯⟩	⟨疒⟩ ⟨㠯⟩	167 10	89.0 0.4

斿	□〈扩〉子	〈扩〉		167	88.0
	□方孓	孓		11	0.4
〈斾〉	□〈扩〉巾	〈扩〉		167	87.0
	□方布	布		12	0.5
旅	□〈扩〉氏	〈扩〉		167	84.2
	□方长	长		15	0.7
〈旅〉	□〈扩〉氏	〈扩〉		167	84.2
	□方长	长		15	0.7
〈旋〉	□〈扩〉〈疋〉	〈扩〉		167	81.5
	□方〈疋〉	〈疋〉		18	1.0
施	□〈扩〉也	〈扩〉		167	42.9
	□方〈㐌〉	〈㐌〉		88	11.9
〈塍〉	□〈朕〉水	〈朕〉		25	92.5
	□〈月/冃/月/月〉⊟〈兴/尖〉水	⊟〈兴/尖〉水		1	0.2
〈勝〉	□〈朕〉力	〈朕〉		25	92.5
	□〈月/冃/月/月〉⊟〈兴/尖〉力	⊟〈兴/尖〉力		1	0.2
Same as above	〈塍〉, 〈滕〉, 〈螣〉				
〈縢〉	□〈朕〉〈衣〉	〈朕〉		25	85.7
	□〈月/冃/月/月〉〈袬〉	〈袬〉		2	0.6
〈縢〉	□〈朕〉〈糸/糸/糸〉	〈朕〉		25	61.0
	□〈月/冃/月/月〉〈縶〉	〈縶〉		7	4.8
〈幐〉	□〈朕〉巾	〈朕〉		25	61.0
	□〈月/冃/月/月〉〈帤〉	〈帤〉		7	4.8
〈幐〉	□〈朕〉巾	〈朕〉		25	61.0
	□〈月/冃/月/月〉〈帤〉	〈帤〉		7	4.8
鰧	□朕鱼	朕		20	90.7
	□〈月〉⊟兴鱼	⊟兴鱼		1	0.3
縢	□朕衣	朕		20	90.7
	□〈月〉⊟兴衣	⊟兴衣		1	0.3
Same as above	螣, 滕				
腦	□朕田	朕		20	82.6
	□〈月〉〈畚〉	〈畚〉		2	0.9

謄	□朕言	朕	20	69.4
	□〈月〉□关言	□关言	4	2.8
〈蟞〉	□〈敖〉虫	〈敖〉	18	89.8
	□男□〈夂/夂〉虫	□〈夂/夂〉虫	1	0.3
〈贅〉	□〈敖〉貝	〈敖〉	18	81.0
	□男□〈夂/夂〉貝	□〈夂/夂〉貝	2	1.1
〈𦞅〉	□〈朕〉足	〈朕〉	15	87.9
	□〈月〉□〈关〉足	□〈关〉足	1	0.4
〈鱻〉	□〈倝〉魚	〈倝〉	15	87.9
	□〈卓〉□〈人/入/ヘ〉魚	□〈人/入/ヘ〉魚	1	0.4
Same as above	〈轤〉, 〈皲〉, 〈㪉〉, 〈鉑〉, 〈戟〉, 〈韓〉, 〈蝪〉, 〈翰〉, 〈榦〉, 〈幹〉, 〈鶾〉, 〈乾〉			
〈幹〉	□〈倝〉干	〈倝〉	15	36.0
	□〈卓〉〈仐〉	〈仐〉	10	16.1
䰃	□倝火	倝	15	87.9
	□〈卓〉□人火	□人火	1	0.4
翰	□倝飛	倝	15	87.9
	□〈卓〉□人飛	□人飛	1	0.4
Same as above	鱻, 鶾, 乾, 蝪, 皲, **翰, 幹, 榦**			
戟	□倝戈	倝	15	77.9
	□〈卓〉□人戈	□人戈	2	1.4
鉑	□倝日	倝	15	77.9
	□〈卓〉□人日	□人日	2	1.4
榦	□倝木	倝	15	77.9
	□〈卓〉□人木	□人木	2	1.4
黱	□朕黑	朕	14	87.1
	□月□关黑	□关黑	1	0.5
鰧	□朕魚	朕	14	87.1
	□月□关魚	□关魚	1	0.5
幐	□朕巾	朕	14	87.1
	□月帣	帣	1	0.5
縢	□朕木	朕	14	76.6
	□月桊	桊	2	1.6

朡	□膄貝	膄	14	54.3
	□［月□癸貝	□癸貝	5	7.0
膝	□膄女	膄	14	54.3
	□［月□癸女	□癸女	5	7.0
塍	□膄土	膄	14	54.3
	□［月□癸土	□癸土	5	7.0
〈隆〉	□〈降〉生	〈降〉	13	86.2
	□〈阝 〉□〈夆〉生	□〈夆〉生	1	0.6
疧	□〈疒〉臣	〈疒〉	11	84.0
	□刊一臣	□一臣	1	0.7
〈癘〉	□〈疒〉〈萬〉	〈疒〉	11	84.0
	□刊一〈萬〉	□一〈萬〉	1	0.7
Same as above	〈癬〉, 〈癉〉, 〈痱〉, 〈疷〉, 〈癱〉			
疾	□〈疒〉矢	〈疒〉	11	71.6
	□刊一矢	□一矢	2	2.4
〈疾〉	□〈疒〉矢	〈疒〉	11	71.6
	□刊一矢	□一矢	2	2.4
Same as above	齎, 〈齎〉			
糵	□穀米	穀	10	82.6
	□嗇□殳米	□殳米	1	0.9
贉	□敄貝	敄	19	81.9
	□矛□攵貝	□攵貝	2	0.9
蝥	□敄虫	敄	19	68.2
	□矛蚤	蚤	4	3.1
傒	□〈伿〉糸	〈伿〉	9	81.0
	□亻支糸	□支糸	1	1.1
傆	□〈伿〉足	〈伿〉	9	81.0
	□亻□支足	□支足	1	1.1
〈傄〉	□〈伿〉月	〈伿〉	9	66.9
	□亻支月	□支月	2	3.4
〈佞〉	□〈仁〉女	〈仁〉	16	79.0
	□〈亻 〉〈妄〉	〈妄〉	2	1.3

俀	□仁女	仁	16	79.0
	□⟨亻⟩⟨妾⟩	⟨妾⟩	2	1.3
⟨临⟩	□⟨𠂤⟩⟨叩/口⟩	⟨𠂤⟩	7	76.6
	□‖⼁𠂤⟨叩/口⟩	□𠂤⟨叩/口⟩	1	1.6
⟨隓⟩	□⟨陸⟩山	⟨陸⟩	7	76.6
	□⟨阝⟩⟨𡍮⟩山	□⟨𡍮⟩山	1	1.6

TABLE 2. □▦亻∣ RB ↔ ▦▦亻∣□ RB (112)

char	structure	component	pn	accuracy
儵	□⟨攸⟩足	⟨攸⟩	54	96.4
	□亻□⟨夂/夊⟩足	□⟨夂/夊⟩足	1	0.1
⟨儠⟩	□⟨攸⟩⟨里⟩	⟨攸⟩	54	96.4
	□亻□⟨夂/夊⟩⟨里⟩	□⟨夂/夊⟩⟨里⟩	1	0.1
Same as above	⟨儵⟩, ⟨脩⟩, ⟨鯈⟩, 鯈, **脩**, ⟨鯈⟩, ⟨倏⟩, ⟨倏⟩, ⟨脩⟩			
⟨脩⟩	□⟨攸⟩貝	⟨攸⟩	54	93.0
	□亻□⟨夂/夊⟩貝	□⟨夂/夊⟩貝	2	0.2
⟨絛⟩	□⟨攸⟩⟨糸 /糸/□糸⟩	⟨攸⟩	54	93.0
	□亻絛	絛	2	0.2
修	□⟨攸⟩彡	⟨攸⟩	54	81.0
	□亻⟨参⟩	⟨参⟩	6	1.1
⟨修⟩	□⟨攸⟩彡	⟨攸⟩	54	81.0
	□亻⟨参⟩	⟨参⟩	6	1.1
⟨𤲃⟩	□⟨攸⟩田	⟨攸⟩	54	59.5
	□亻⟨备⟩	⟨备⟩	16	5.3
⟨條⟩	□⟨攸⟩⟨朩/木⟩	⟨攸⟩	54	47.9
	□亻⟨条⟩	⟨条⟩	24	9.5
倏	□攸足	攸	36	94.7
	□亻□夊足	□夊足	1	0.1
絛	□攸糸	攸	36	94.7
	□亻絛	絛	1	0.1
Same as above	**鯈, 鱐, 條**, 倏, 脩, 儵			

char	structure	component	pn	accuracy
傆	□攴貝 □〈亻〉□攵貝	攴 □攵貝	36 2	89.8 0.3
倏	□攴火 □〈亻〉□攵火	攴 □攵火	9 9	25.0 25.1
俻	□攴田 □〈亻〉备	攴 备	9 13	16.7 35.0

TABLE 3. □□ ATR ↔ □ A □ TR if T is *tare* (121)

char	structure	component	pn	accuracy
庴	□广昝 □〈卜〉□厂昝	广 □厂昝	65 1	97.0 0.1
庌	□广子 □〈卜〉□厂子	广 □厂子	65 1	97.0 0.1
Same as above	〈庴〉, 〈庩〉, 〈庱〉, 〈庴〉			
庋	□广又 □〈卜〉□厂又	广 □厂又	65 2	94.1 0.1
〈庋〉	□广又 □〈卜〉□厂又	广 □厂又	65 2	94.1 0.1
麳	□〈𡤼〉來 □枚㣇	〈𡤼〉 㣇	26 1	92.7 0.2
〈贅〉	□〈𡤼〉貝 □枚〈𧵐〉	〈𡤼〉 〈𧵐〉	26 1	92.7 0.2
Same as above	〈麳〉, 〈鏊〉, 〈�〉, 〈麳〉, 〈㣇〉			
〈鶩〉	□〈𡤼〉鳥 □枚鴈	〈𡤼〉 鴈	26 2	86.2 0.6
〈麳〉	□〈𡤼〉〈來〉 □枚〈㣇〉	〈𡤼〉 〈㣇〉	26 2	86.2 0.6
Same as above	〈�〉, 〈㣇〉, 〈㣇〉			
〈幣〉	□〈𡤼〉巾 □枚〈帀〉	〈𡤼〉 〈帀〉	26 3	80.4 1.1

〈氅〉	▣〈产〉文 ▤枚〈反〉	〈产〉 〈反〉	26 4	75.1 1.8
縠	▣〈产〉水 ▤枚〈床〉	〈产〉 〈床〉	26 4	75.1 1.8
Same as above	〈縠〉, 〈氅〉, 〈勞〉			
〈氅〉	▣〈产〉干 ▤枚斤	〈产〉 斤	26 5	70.3 2.6
〈嫠〉	▣〈产〉女 ▤枚▣厂女	〈产〉 ▣厂女	26 5	70.3 2.6
釐	▣〈产〉里 ▤枚厘	〈产〉 厘	26 6	66.0 3.6
〈犛〉	▣〈产〉牛 ▤枚▣厂牛	〈产〉 ▣厂牛	26 6	66.0 3.6
〈氂〉	▣〈产〉毛 ▤枚▣厂毛	〈产〉 ▣厂毛	26 7	62.1 4.5
〈勞〉	▣〈产〉万 ▤枚厉	〈产〉 厉	26 13	44.4 11.2
〈釐〉	▣〈产〉〈里〉 ▤枚〈厘〉	〈产〉 〈厘〉	26 44	13.8 39.6
〈斎〉	▣〈产〉初 ▤〈立/文〉▣厂初	〈产〉 ▣厂初	22 1	91.5 0.2
〈廉〉	▣〈产〉〈兼〉 ▤〈立/文〉〈廉〉	〈产〉 〈廉〉	22 1	91.5 0.2
Same as above	〈炭〉, 〈序〉, 〈廉〉			
〈廉〉	▣〈产〉兼 ▤〈立/文〉▣厂兼	〈产〉 ▣厂兼	22 2	84.0 0.7
〈詹〉	▣〈产〉言 ▤〈立/文〉▣厂言	〈产〉 ▣厂言	22 6	61.7 4.6
〈炭〉	▣〈产〉火 ▤〈立/文〉灰	〈产〉 灰	22 24	22.9 27.3
〈店〉	▣〈产〉舌 ▤〈刀/ク/力/ク/ク〉▣厂舌	〈产〉 ▣厂舌	21 1	91.1 0.2

〈詹〉	□〈产〉言 □〈刀/冖/力/ク/ク〉□ 厂言	〈产〉 □厂言	21 1	91.1 0.2
Same as above	〈厃〉, 〈危〉			
〈佚〉	□〈产〉失 □〈刀/冖/力/ク/ク〉□ 厂失	〈产〉 □厂失	21 2	83.4 0.8
〈佚〉	□〈产〉矢 □〈刀/冖/力/ク/ク〉厌	〈产〉 厌	21 7	56.3 6.3
危	□〈产〉〈巳〉 □〈刀/冖/力/ク/ク〉厄	〈产〉 厄	21 42	11.1 44.5
庿	□广舌 □〈ク〉□厂舌	广 □厂舌	21 1	91.1 0.2
詹	□广言 □〈ク〉□厂言	广 □厂言	21 1	91.1 0.2
佚	□广失 □〈ク〉□厂失	广 □厂失	21 2	83.4 0.8
佚	□广矢 □〈ク〉厌	广 厌	21 7	56.3 6.3
〈巌〉	□〈屵〉〈磊〉 □山□〈厂/丁〉〈磊〉	〈屵〉 □〈厂/丁〉〈磊〉	14 1	87.1 0.5
〈巎〉	□〈屵〉〈堯〉 □山□〈厂/丁〉〈堯〉	〈屵〉 □〈厂/丁〉〈堯〉	14 1	87.1 0.5
Same as above	〈㠚〉, 〈嵞〉, 〈巑〉, 〈嶜〉, 〈嵪〉, 〈㞳〉, 〈崶〉, 〈崖〉, 〈㟅〉, 〈嵓〉, 崔			
㦱	□产〈夂〉 □立□厂〈夂〉	产 □厂〈夂〉	11 1	84.0 0.7
廉	□产兼 □立□厂兼	产 □厂兼	11 1	84.0 0.7
廉	□产兼 □立□厂兼	产 □厂兼	11 2	71.6 2.4
彦	□产彡 □立彦	产 彦	11 5	47.3 9.8

| 彦 | □产言 | 产 | 11 | 41.9 |
| | □立□厂言 | □厂言 | 6 | 12.5 |

TABLE 4. □□ ADR ↔ □ A □ DR if D is not *tare* (122)

char	structure	component	p*n*	accuracy
死	□〈歹〉巳	〈歹〉	451	99.6
	□一□〈夕/攵/夂〉巳	□〈夕/攵/夂〉巳	1	0.1
〈死〉	□〈歹〉巳/己/巳/㔾	〈歹〉	451	99.6
	□一□〈夕/攵/夂〉〈巳/己/巳/㔾〉	□〈夕/攵/夂〉〈巳/己/巳/㔾〉	1	0.1
〈死〉	□〈歹〉《匕》/七	〈歹〉	451	99.6
	□一□〈夕/攵/夂〉《匕》/七	□〈夕/攵/夂〉《匕》/七	1	0.1
死	□歹匕	歹	113	93.3
	□一歼	歼	4	0.2
皀	□皀卩	皀	33	94.2
	□白印	印	1	0.1
〈皀〉	□〈皀〉卩	〈皀〉	23	91.8
	□白□〈匕〉卩	□〈匕〉卩	1	0.2
石	□〈丆〉口	〈丆〉	22	91.5
	□一□丿口	□丿口	1	0.2
〈布〉	□〈丆〉巾	〈丆〉	22	91.5
	□一□丿巾	□丿巾	1	0.2
㝱	□宀蔓	宀	19	90.3
	□宀□爿蔓	□爿蔓	1	0.3
寀	□宀釆	宀	19	90.3
	□宀〈釆〉	〈釆〉	1	0.3
Same as above	寢, 癗, 寙, 寀, 癁, 㝱, 寙, 〈寢〉			
寢	□宀夒	宀	19	81.9
	□宀〈㝱〉	〈㝱〉	2	0.9
㝴	□宀臬	宀	19	74.6
	□宀□爿臬	□爿臬	3	1.9
寤	□宀吾	宀	19	74.6
	□宀〈悟〉	〈悟〉	3	1.9

char	structure	component		
病	□〈宀〉丙 □⺧□〈爿/丬〉丙	〈宀〉 □〈爿/丬〉丙	17 1	89.2 0.3
〈寢〉	□〈宀〉㝱 □⺧□〈爿/丬〉㝱	〈宀〉 □〈爿/丬〉㝱	17 1	89.2 0.3
Same as above	〈癗〉,〈癮〉,〈寐〉,〈癌〉,〈癯〉,〈瘖〉,〈寐〉,〈病〉,〈癆〉,〈寐〉			
寤	□〈宀〉言 □⺧□〈爿/丬〉言	〈宀〉 □〈爿/丬〉言	17 2	80.1 1.1
〈寤〉	□〈宀〉言 □⺧□〈爿/丬〉言	〈宀〉 □〈爿/丬〉言	17 2	80.1 1.1
着	□羊目 □〈⺷〉□丿目	羊 □丿目	12 1	85.2 0.6
羞	□羊丑 □〈⺷〉□丿丑	羊 □丿丑	12 1	85.2 0.6
Same as above	〈鱃〉,〈韹〉,〈詹〉			

TABLE 5. □□ E 彳 R ↔ □ E □彳 R (131)

char	structure	component	p*n*	accuracy
㢮	□尸曳 □尸□彳曳	尸 □彳曳	9 1	81.0 1.1
屡	□尸婁 □尸□彳婁	尸 □彳婁	9 1	81.0 1.1
屩	□尸喬 □尸僑	尸 僑	9 1	81.0 1.1
〈㢮〉	□尸〈曳/曳〉 □尸□彳〈曳/曳〉	尸 □彳〈曳/曳〉	9 1	81.0 1.1
〈屡〉	□尸〈婁/婁/婁〉 □尸□彳〈婁/婁/婁〉	尸 □彳〈婁/婁/婁〉	9 1	81.0 1.1
屟	□尸枼 □尸□彳枼	尸 □彳枼	9 1	81.0 1.1
〈屢〉	□尸娄 □尸□彳娄	尸 □彳娄	9 1	81.0 1.1

	□尸世	尸世	9	81.0
屜	□尸□彳世	□彳世	1	1.1
〈屩〉	□尸〈喬〉	尸	9	66.9
	□尸〈僑〉	〈僑〉	2	3.4

TABLE 6. □□ EAB ↔ □ E □ AB if A = non 彳 (132)

char	structure	component	pn	accuracy
屟	□尾水	尾水	145	98.6
	□尸□毛水	□毛水	1	0.1
〈麜〉	□〈麻〉仌	〈麻〉	133	98.5
	□广□〈林/朩〉仌	□〈林/朩〉仌	1	0.1
〈麜〉	□〈麻〉木	〈麻〉	133	98.5
	□广□〈林/朩〉木	□〈林/朩〉木	1	0.1
Same as above	〈庵〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈魔〉, 〈麿〉, 〈麿〉, 〈麿〉, 〈摩〉, 〈麿〉, 〈麿〉			
〈屎〉	□〈尼〉木	〈尼〉	97	98.0
	□尸□〈匕〉木	□〈匕〉木	1	0.1
麼	□麻多	麻	77	97.5
	□广□〈林〉多	□〈林〉多	1	0.1
糜	□麻米	麻	77	97.5
	□广□〈林〉米	□〈林〉米	1	0.1
Same as above	賡, 麿, 麿, 麿, 庵, 麿, 麿, 麿, 麿, 麿, 麿, 麿, 麿, 麿, 糜, 麿, 麿, 麿, 麿, 麿, 麿, 麿, 麿, 〈麿〉, 〈麿〉, 〈麿〉			
膚	□麻日	麻	77	95.0
	□广賮	賮	2	0.1
麼	□麻女	麻	77	95.0
	□广□〈林〉女	□〈林〉女	2	0.1
靡	□麻非	麻	77	95.0
	□广□〈林〉非	□〈林〉非	2	0.1
〈鷹〉	□〈厥〉〈鸟〉	〈厥〉	67	97.1
	□〈厂/丆〉□〈欹〉〈鸟〉	□〈欹〉〈鸟〉	1	0.1
〈腐〉	□〈厥〉肉	〈厥〉	67	97.1
	□〈厂/丆〉□〈欹〉肉	□〈欹〉肉	1	0.1
〈鷹〉	□〈厥〉鳥	〈厥〉	67	97.1
	□〈厂/丆〉□〈欹〉鳥	□〈欹〉鳥	1	0.1

〈蠤〉	□〈厥〉虫 □〈厂/丆〉□〈欮〉虫	〈厥〉 □〈欮〉虫	67 2	94.3 0.1
〈�装〉	□康心 □广□隶心	康 □隶心	55 1	96.5 0.1
〈庑〉	□庚〈凡〉 □广□夬〈凡〉	庚 □夬〈凡〉	51 1	96.2 0.1
〈庳〉	□庚十 □广□夬十	庚 □夬十	51 1	96.2 0.1
Same as above	〈庼〉, 〈庼〉, 〈庼〉, 廙			
〈厡〉	□〈原〉水 □〈厂/丆〉□泉水	〈原〉 □泉水	48 1	96.0 0.1
鷹	□麻鳥 □广〈鵥〉	麻 〈鵥〉	45 1	95.7 0.1
麻	□麻口 □广〈㗊〉	麻 〈㗊〉	45 1	95.7 0.1
Same as above	麋, 麇, 摩, 魔, 磨, 魔, 麂, 麼, 摩			
譍	□麻言 □广〈誉〉	麻 〈誉〉	45 2	91.7 0.2
麿	□麻呂 □广□林呂	麻 □林呂	45 3	87.9 0.4
庽	□庐匃 □广□屮匃	庐 □屮匃	43 1	95.5 0.1
褒	□庐衣 □广□屮衣	庐 □屮衣	43 1	95.5 0.1
Same as above	〈賡〉, 〈廥〉, 〈廥〉, 〈廥〉, 〈庽〉, 〈庽〉, 庽			
庺	□庐且 □广□屮且	庐 □屮且	43 2	91.3 0.2
〈庶〉	□庐灬 □广□屮灬	庐 □屮灬	43 2	91.3 0.2
Same as above	〈庼〉, 〈庺〉			
庲	□庐厽 □广□屮厽	庐 □屮厽	43 3	87.4 0.5

鹿	□声比 □广□出比	声 □出比	43 3	87.4 0.5
〈庶〉	□声厶 □广□出厶	声 □出厶	43 3	87.4 0.5
〈鹿〉	□声〈比〉 □广□出〈比〉	声 □出〈比〉	43 5	80.3 1.1
〈麤〉	□〈厭〉食 □厂〈饕〉	〈厭〉 〈饕〉	35 1	94.5 0.1
〈麤〉	□〈厭〉面 □厂〈靨〉	〈厭〉 〈靨〉	35 1	94.5 0.1
Same as above	麚,〈麘〉,麇,〈麛〉			
歷	□〈厤〉止 □〈厂／ア〉□秝止	〈厤〉 □秝止	34 1	94.4 0.1
〈曆〉	□〈厤〉田 □〈厂／ア〉〈畨〉	〈厤〉 〈畨〉	34 1	94.4 0.1
Same as above	〈歴〉,〈鸁〉,〈麠〉,〈曆〉,〈麜〉,〈麧〉,〈壓〉,〈歷〉,〈曆〉,〈麚〉			
〈曆〉	□〈厤〉甘 □〈厂／ア〉〈畨〉	〈厤〉 〈畨〉	34 2	89.2 0.3
磨	□〈厤〉〈石〉 □〈厂／ア〉□秝〈石〉	〈厤〉 □秝〈石〉	34 2	89.2 0.3
〈歷〉	□〈厤〉〈止〉 □〈厂／ア〉□秝〈止〉	〈厤〉 □秝〈止〉	34 2	89.2 0.3
膺	□〈雁〉目 □广□〈隹〉目	〈雁〉 □〈隹〉目	31 1	93.8 0.1
鷹	□〈雁〉口 □广□〈隹〉口	〈雁〉 □〈隹〉口	31 1	93.8 0.1
Same as above	鷹,膺,膺,鷹,麐,〈鷹〉			
應	□〈雁〉心 □广□〈隹〉心	〈雁〉 □〈隹〉心	31 2	88.2 0.4
鷹	□〈雁〉鳥 □广□〈隹〉鳥	〈雁〉 □〈隹〉鳥	31 2	88.2 0.4
Same as above	膺,膺			

廞	□府衣 □广□付衣	府 □付衣	29 1	93.4 0.2
腐	□府肉 □广□付肉	府 □付肉	29 1	93.4 0.2
〈賔〉	□府貝 □广□付貝	府 □付貝	29 2	87.5 0.5
庡	□府天 □广□付天	府 □付天	29 3	82.1 0.9
〈庲〉	□府天 □广□付天	府 □付天	29 3	82.1 0.9
厴	□厭甲 □厂□猒甲	厭 □猒甲	27 1	93.0 0.2
庍	□灰匕 □厂□火匕	灰 □火匕	24 1	92.2 0.2
〈厐〉	□灰〈匕〉 □厂□火〈匕〉	灰 □火〈匕〉	24 1	92.2 0.2
〈愿〉	□厚心 □厂□〈旱〉心	厚 □〈旱〉心	21 1	91.1 0.2
厄	□〈厂〉〈巳〉 □〈厂〉□一〈巳〉	〈厂〉 □一〈巳〉	20 1	90.7 0.3
〈厄〉	□〈厂〉〈㔾〉 □〈厂〉□一〈㔾〉	〈厂〉 □一〈㔾〉	20 1	90.7 0.3
〈厄〉	□〈厂〉〈已/己/巳/㔾〉 □〈厂〉□一〈已/己/巳/㔾〉	〈厂〉 □一〈已/己/巳/㔾〉	20 1	90.7 0.3
〈斤〉	□〈厂〉丁 □〈厂〉〈亍〉	〈厂〉 〈亍〉	20 2	82.6 0.9
〈厃〉	□〈厂〉乙 □〈厂〉〈乙〉	〈厂〉 〈乙〉	20 2	82.6 0.9
厄	□〈厂〉巴 □〈厂〉□一巴	〈厂〉 □一巴	20 2	82.6 0.9
后	□〈厂〉口 □〈厂〉〈司〉	〈厂〉 〈司〉	20 68	5.2 59.8
〈帀〉	□〈厂〉巾 □〈厂〉帀	〈厂〉 帀	20 80	4.0 64.0
犀	□尿牛 □尸□水牛	尿 □水牛	16 1	88.6 0.4
〈廲〉	□庫〈蚰〉 □广□車〈蚰〉	庫 □車〈蚰〉	16 1	88.6 0.4
麤	□厤鬲 □厂□秝鬲	厤 □秝鬲	16 1	88.6 0.4

曆	□厤日	厤	16	88.6
	□厂⊟秝日	⊟秝日	1	0.4
Same as above	曆, 曆, 曆, 曆, 曆, 曆, 曆, 魘, 歴			
厬	□厤火	厤	16	79.0
	□厂⊟秝火	⊟秝火	2	1.3
曆	□厤心	厤	16	79.0
	□厂⊟秝心	⊟秝心	2	1.3
曆	□厥骨	厥	16	88.6
	□厂⊟欮骨	⊟欮骨	1	0.4
蠜	□厥虫	厥	16	88.6
	□厂⊟欮虫	⊟欮虫	1	0.4
〈鴈〉	□雁〈直〉	雁	10	82.6
	□厂⊟〈隹〉〈直〉	⊟〈隹〉〈直〉	1	0.9
〈贋〉	□雁〈貝〉	雁	10	82.6
	□厂⊟〈隹〉〈貝〉	⊟〈隹〉〈貝〉	1	0.9
贋	□雁貝	雁	10	69.4
	□厂⊟〈隹〉貝	⊟〈隹〉貝	2	2.8
〈㮇〉	□〈厌〉木	〈厌〉	10	82.6
	□厂⊟犬木	⊟犬木	1	0.9
〈㧑〉	□〈厌〉手	〈厌〉	10	82.6
	□厂⊟犬手	⊟犬手	1	0.9
Same as above	〈㤪〉, 〈㤫〉			
〈㙑〉	□〈厌〉土	〈厌〉	10	69.4
	□厂⊟犬土	⊟犬土	2	2.8
庶	□庻灬	庻	9	81.0
	□广㶳	㶳	1	1.1
度	□庻又	庻	9	81.0
	□广㠯	㠯	1	1.1
Same as above	廈, 庶			
庹	□庻火	庻	9	56.3
	□广㶳	㶳	3	6.3
席	□庻巾	庻	9	47.9
	□广帯	帯	4	9.5
屬	□犀〈蜀〉	犀	9	81.0
	□尸⊟牛〈蜀〉	⊟牛〈蜀〉	1	1.1

char	structure	component	pn	accuracy
犀	□犀**牛** / □尸□丰**牛**	犀 / □丰**牛**	9 / 1	81.0 / 1.1
Same as above	犀,〈屎〉,〈犀〉,〈屆〉,〈犀〉			
愿	□**原心** / □厂□泉**心**	**原** / □泉**心**	8 / 1	79.0 / 1.3
〈贋〉	□**鴈真** / □厂□〈倠〉**真**	**鴈** / □〈倠〉**真**	7 / 1	76.6 / 1.6
〈鷹〉	□**鴈火** / □厂□〈倠〉**火**	**鴈** / □〈倠〉**火**	7 / 1	76.6 / 1.6
〈庶〉	□庐灬 / □广□〈卝/卝/**廿**/卝/丱〉灬	庐 / □〈卝/卝/**廿**/卝/丱〉灬	7 / 1	76.6 / 1.6
〈度〉	□庐**又** / □广□〈卝/卝/**廿**/卝/丱〉**又**	庐 / □〈卝/卝/**廿**/卝/丱〉**又**	7 / 1	76.6 / 1.6
Same as above	〈庶〉,〈廐〉			
〈席〉	□庐**巾** / □广〈帯/帯/帯/帯〉	庐 / 〈帯/帯/帯/帯〉	7 / 3	49.0 / 9.1
〈庹〉	□庐**火** / □广〈芡〉	庐 / 〈芡〉	7 / 5	34.0 / 17.4

TABLE 7. □□□ LRB ↔ □□□ LBR (210)

char	structure	component	pn	accuracy
〈穀〉	□〈殻〉**赤** / □□□吉**赤**〈殳〉	〈殻〉 / □吉**赤**	40 / 1	95.2 / 0.1
〈穀〉	□〈殻〉**羊** / □□□吉**羊**〈殳〉	〈殻〉 / □吉**羊**	40 / 1	95.2 / 0.1
Same as above	〈穀〉,〈穀〉,〈穀〉,〈穀〉,穀,**穀**,穀,〈穀〉,〈穀〉			
〈豰〉	□〈殻〉〈豕〉 / □□□吉〈豕〉〈殳〉	〈殻〉 / □吉〈豕〉	40 / 2	90.7 / 0.3
穀	□〈殻〉**木** / □□素〈殳〉	〈殻〉 / 素	40 / 2	90.7 / 0.3
Same as above	**穀**,〈穀〉,〈穀〉,〈穀〉,〈穀〉,〈穀〉,〈穀〉			

〈㲉〉	□〈㲉〉出 □日壴出〈殳〉	〈㲉〉 日壴出	40 3	86.5 0.5
〈㲋〉	□〈㲉〉子 □亭〈殳〉	〈㲉〉 亭	40 3	86.5 0.5
Same as above	〈㲊〉, 〈㲉〉			
〈㲉〉	□〈㲉〉卵 □日壴卵〈殳〉	〈㲉〉 日壴卵	40 4	82.6 0.9
㲉	□㲉缶 □日壴缶殳	㲉 日壴缶	32 1	94.0 0.1
㲉	□㲉口 □壴殳	㲉 壴	32 1	94.0 0.1
Same as above	㲉, 㲉, 㲉, 㲉			
㲉	□㲉林 □日壴林殳	㲉 日壴林	32 2	88.6 0.4
㲉	□㲉目 □日壴目殳	㲉 日壴目	32 2	88.6 0.4
Same as above	㲉, 㲉, 㲉, 㲉			
㲉	□㲉子 □亭殳	㲉 亭	32 3	83.6 0.8
㲉	□㲉火 □日壴火殳	㲉 日壴火	32 3	83.6 0.8
㲉	□㲉出 □日壴出殳	㲉 日壴出	32 3	83.6 0.8
㲉	□㲉卵 □日壴卵殳	㲉 日壴卵	32 4	79.0 1.3
穀	□㲉禾 □〈彙〉殳	㲉 〈彙〉	32 4	79.0 1.3
〈頯〉	□〈須〉女 □日彡女〈頁/頁〉	〈須〉 日彡女	28 1	93.2 0.2
〈穎〉	□〈頃〉〈耒〉 □日〈匕/七〉〈耒〉〈頁/頁〉	〈頃〉 日〈匕/七〉〈耒〉	20 1	90.7 0.3
〈穎〉	□〈頃〉禾 □日〈匕/七〉禾〈頁/頁〉	〈頃〉 日〈匕/七〉禾	20 2	82.6 0.9
〈頴〉	□〈頃〉示 □〈祟〉〈頁/頁〉	〈頃〉 〈祟〉	20 8	51.0 8.2

char	structure	component	pn	accuracy
〈縠〉	□□□士宀〈殳〉〈孚〉 □□□士宀〈孚〉〈殳〉	□□士宀〈殳〉 □□士宀〈孚〉	11 1	84.0 0.7
〈縠〉	□□□士宀〈殳〉士 □□□士宀士〈殳〉	□□士宀〈殳〉 □□士宀士	11 1	84.0 0.7
Same as above	〈縠〉, 〈縠〉, 〈縠〉, 〈縠〉, 〈縠〉, 〈縠〉			
〈穀〉	□□□士宀〈殳〉□口米 □□□士宀口米〈殳〉	□□士宀〈殳〉 □□士宀口米	11 2	71.6 2.4
〈穀〉	□□□士宀〈殳〉告 □□□士宀告〈殳〉	□□士宀〈殳〉 □□士宀告	11 2	71.6 2.4
〈穀〉	□□□士宀〈殳〉牛 □□□士宀牛〈殳〉	□□士宀〈殳〉 □□士宀牛	11 2	71.6 2.4
穀	□殻告 □□声告殳	殻 □声告	10 1	82.6 0.9
穀	□殻牛 □□声牛殳	殻 □声牛	10 1	82.6 0.9
Same as above	殻, 穀, 穀, 穀, 穀, **穀**			
叡	□〈叔〉谷 □□□〈卢〉谷又	〈叔〉 □〈卢〉谷	7 1	76.6 1.6
〈叡〉	□〈叔〉〈谷/谷〉 □□□〈卢〉〈谷/谷〉又	〈叔〉 □〈卢〉〈谷/谷〉	7 1	76.6 1.6
叡	□〈叔〉〈贝〉 □□□〈卢〉〈贝〉又	〈叔〉 □〈卢〉〈贝〉	7 2	60.5 5.0
〈叡〉	□〈叔〉〈贝〉 □□□〈卢〉〈贝〉又	〈叔〉 □〈卢〉〈贝〉	7 2	60.5 5.0
〈叡〉	□〈叔〉贝 □□□〈卢〉贝又	〈叔〉 □〈卢〉贝	7 2	60.5 5.0
〈叡〉	□〈台〉立 □□□台立彐	〈台〉 □台立	7 1	76.6 1.6

TABLE 8. □□ AEM ↔ □ A □ EM if E is *kamae* (411)

char	structure	component	pn	accuracy
〈卤〉	□占〈囗〉 □〈卜〉□口〈囗〉	占 □口〈囗〉	199 1	99.0 0.1
鹵	□占〈※〉 □〈卜〉図	占 図	199 18	84.1 0.7

char	structure	component	p*n*	accuracy
〈卤〉	▣占〈乂〉	占	199	69.3
	▣〈卜〉〈囟〉	〈囟〉	40	2.9
〈互〉	▣〈亙〉丶	〈亙〉	58	96.6
	▣一▣〈亙〉丶	▣〈亙〉丶	1	0.1
〈囱〉	▣白〈夕/夂/夊〉	白	33	94.2
	▣丿▣口〈夕/夂/夊〉	▣口〈夕/夂/夊〉	1	0.1
囱	▣白夊	白	33	79.5
	▣丿囟	囟	4	1.2
〈囪〉	▣白夊	白	33	79.5
	▣丿囟	囟	4	1.2
囟	▣白〈乂〉	白	33	20.4
	▣丿〈囟〉	〈囟〉	40	30.1

TABLE 9. ▣▣ ABM ↔ ▣ A ▣ MB if both A and B are not enclosure (414)

char	structure	component	p*n*	accuracy
〈袤〉	▣〈衣〉〈丑〉	〈衣〉	389	99.5
	▣〈宀〉▣〈丑〉〈𠂇〉	▣〈丑〉〈𠂇〉	1	0.1
〈褭〉	▣〈衣〉〈嬰〉	〈衣〉	389	99.5
	▣〈宀〉▣〈嬰〉〈𠂇〉	▣〈嬰〉〈𠂇〉	1	0.1
Same as above	〈襄〉, 〈裒〉, 〈褺〉, 〈裛〉, 〈褱〉, 〈襞〉, 〈襪〉, 〈裒〉, 〈褋〉, 〈裏〉, 〈褻〉, 〈裛〉, 〈褒〉, 〈褒〉, 〈褻〉, 〈褒〉, 〈衯〉, 〈裒〉, 〈褢〉, 〈褒〉, 〈褋〉, 〈褖〉, 〈裒〉, 〈裛〉, 〈裏〉, 〈袠〉, 〈裵〉, 〈衷〉, 〈褏〉, 〈裒〉, 〈褒〉, 〈衷〉, 〈裒〉, 〈褧〉, 〈褎〉, 〈裒〉, 〈衷〉, 〈表〉, 〈衮〉, 〈袤〉, 〈褒〉, 〈表〉, 〈裒〉, 〈裒〉, 〈褒〉, 〈褺〉, 裒, 〈裛〉, 〈裛〉, 〈褒〉, 〈褏〉, 〈褧〉, 襄, 〈褒〉, 裒, 〈袞〉, 〈裏〉, 〈裒〉, 〈裒〉, 表, 袞, 〈裒〉, 〈裒〉, 〈衷〉, 〈裒〉, 〈袞〉, 〈褻〉, 〈裒〉			
〈襄〉	▣〈衣〉〈珏〉	〈衣〉	389	99.0
	▣〈宀〉羧	羧	2	0.1
〈襄〉	▣〈衣〉罪	〈衣〉	389	99.0
	▣〈宀〉▣罪〈𠂇〉	▣罪〈𠂇〉	2	0.1
Same as above	〈哀〉, 〈襄〉			
褺	▣衣執	衣	185	98.9
	▣亠〈埶〉	〈埶〉	1	0.1
袞	▣衣火	衣	185	98.9
	▣亠〈炎〉	〈炎〉	1	0.1

Same as above above	亵, 褻, 裘, 褭, 衮, 衺, 裛, 裛, 衮, 袠, 褩, 袠, 裵, 衺, 䙊, 裛, 衺, 衺, 裵, 褻, 褸, 㦻, 衮, 袠, 袠, 袠, **褺**, **褽**, **袠**, **裏**, **褽**, **衷**, **䙂**, **裛**, **韢**, **褺**, **褻**, **韢**, **褻**, **褧**, **褰**, **褺**, **褒**, **裵**, **裛**, **衺**, **衺**, **衷**, **衺**, **袠**, **裵**, **褱**, **裏**, **裘**			
襄	▢衣𡈼 ▢亠▢𡈼仅	衣 ▢𡈼仅	185 2	97.9 0.1
襄	▢衣罪 ▢亠▢罪仅	衣 ▢罪仅	185 2	97.9 0.1
Same as above	裹, 褢			
哀	▢衣口 ▢亠良	衣 良	185 7	92.8 0.2
〈啻〉	▢音中 ▢立㕭	音 㕭	153 1	98.7 0.1
〈侖〉	▢〈侖〉〈⺍〉 ▢〈入/𠆢/个/亼/厶〉▢〈⺍〉〈冊/**冊**/**冊**〉	〈侖〉 ▢〈⺍〉〈冊/**冊**/**冊**〉	83 1	97.6 0.1
〈亜〉	▢〈二〉▥口勹口 ▢一▢▥口勹口一	〈二〉 ▢▥口勹口一	31 1	93.8 0.1
〈舟〉	▢〈二〉舟 ▢一▢舟一	〈二〉 ▢舟一	31 1	93.8 0.1
冋	▢〈二〉冋 ▢一▢冋一	〈二〉 ▢冋一	31 2	88.2 0.4
〈亙〉	▢〈二〉㫃 ▢一▢㫃一	〈二〉 ▢㫃一	31 2	88.2 0.4
〈囘〉	▢〈二〉回 ▢一▢回一	〈二〉 ▢回一	31 2	88.2 0.4
橐	▢橐包 ▢严▢包木	橐 ▢包木	21 1	91.1 0.2
橐	▢橐𣪊 ▢严▢𣪊木	橐 ▢𣪊木	21 1	91.1 0.2
Same as above	橐, 橐, 〈橐〉, 〈橐〉, 〈橐〉, 〈橐〉, 〈橐〉, 〈橐〉, 〈橐〉, 〈橐〉			
橐	▢橐〈皿〉 ▢严▢〈皿〉木	橐 ▢〈皿〉木	21 2	83.4 0.8
〈橐〉	▢橐〈皿〉 ▢严▢〈皿〉木	橐 ▢〈皿〉木	21 2	83.4 0.8
橐	▢橐〈臼〉 ▢严〈梟〉	橐 〈梟〉	21 3	76.6 1.6

char	structure	component	pn	accuracy
橐	⿴橐〈石〉 ⿱严〈橐〉	橐 〈橐〉	21 9	49.0 9.1
〈橐〉	⿴橐非 ⿱严棐	橐 棐	21 13	38.1 14.7
藜	⿱艹黎 ⿱〈艹〉⿱黎艹	艹 ⿱黎艹	17 1	89.2 0.3
荞	⿱艹大 ⿱〈艹〉⿱大艹	艹 ⿱大艹	17 1	89.2 0.3
Same as above	蘽, 薻, 蕠, 舊, 菲, 蘼, 葺, 葬, 茶, 茸, 〈葺〉			
囲	⿴〈囲〉桐 ⿱⿰田田⿰桐⿰田田	〈囲〉 ⿱桐⿰田田	12 1	85.2 0.6
〈囲〉	⿴〈囲〉〈隹/缶/舌〉 ⿱⿰田田⿰〈隹/缶/舌〉⿰田田	〈囲〉 ⿱〈隹/缶/舌〉⿰田田	12 1	85.2 0.6
Same as above	囲, 〈囲〉, 〈囲〉			
囲	⿴〈囲〉辰 ⿱⿰田田⿰辰⿰田田	〈囲〉 ⿱辰⿰田田	12 2	73.5 2.1
〈囲〉	⿴〈囲〉辰 ⿱⿰田田⿰辰⿰田田	〈囲〉 ⿱辰⿰田田	12 2	73.5 2.1
囲	⿴囲缶 ⿱⿰田田⿰缶⿰田田	囲 ⿱缶⿰田田	12 1	85.2 0.6
〈囲〉	⿴〈品〉莫 ⿱〈叩〉⿱莫〈叩〉	〈品〉 ⿱莫〈叩〉	11 1	84.0 0.7
〈品〉	⿴〈品〉⿰太太 ⿱〈叩〉⿱⿰太太〈叩〉	〈品〉 ⿱⿰太太〈叩〉	11 1	84.0 0.7

TABLE 10. ⿰⿱⿰ LRBA -> ⿱⿲ LARB (511)

char	structure	component	pn	accuracy
〈鬻〉	⿴〈鬲〉〈咸〉 ⿱⿲弓〈咸〉弓〈鬲〉	〈鬲〉 ⿲弓〈咸〉弓	43 1	95.5 0.1
〈鬻〉	⿴〈鬲〉〈去〉 ⿱⿲弓〈去〉弓〈鬲〉	〈鬲〉 ⿲弓〈去〉弓	43 1	95.5 0.1
Same as above	〈鬻〉, 〈鬻〉, 〈鬻〉, 〈鬻〉, 鬻, 〈鬻〉			

〈彌〉〈鬲〉	▢〈鬻〉束 ▢〈粥〉〈鬲〉	〈鬻〉 〈粥〉	43 2	91.3 0.2
〈鬻〉〈鬲〉	▢〈鬻〉米 ▢▢弓米弓〈鬲〉	〈鬻〉 ▥弓米弓	43 2	91.3 0.2
Same as above	〈鬻〉, 〈鬻〉, 〈鬻〉, 〈鬻〉, 〈鬻〉, 〈鬻〉, 〈鬻〉, 〈鬻〉			
〈彌〉〈鬲〉	▢〈鬻〉〈匆〉 ▢〈粥〉〈鬲〉	〈鬻〉 〈粥〉	43 3	87.4 0.5
〈彌〉〈鬲〉	▢〈鬻〉〈叁〉 ▢▢弓〈叁〉弓〈鬲〉	〈鬻〉 ▥弓〈叁〉弓	43 3	87.4 0.5
〈彌〉〈鬲〉	▢〈鬻〉每 ▢▢弓每弓〈鬲〉	〈鬻〉 ▥弓每弓	43 3	87.4 0.5
〈彌〉〈鬲〉	▢〈鬻〉孝 ▢弓孝弓〈鬲〉	〈鬻〉 ▥弓孝弓	43 3	87.4 0.5
彌鬲	▢鬻曾 ▢▢弓曾弓鬲	鬻 ▥弓曾弓	19 1	90.3 0.3
彌鬲	▢鬻辱 ▢▢弓辱弓鬲	鬻 ▥弓辱弓	19 1	90.3 0.3
Same as above	鬻, 鬻, 鬻, 鬻			
彌鬲	▢鬻弗 ▢▢弓弗弓鬲	鬻 ▥弓弗弓	19 2	81.9 0.9
彌鬲	▢鬻酉 ▢▢弓酉弓鬲	鬻 ▥弓酉弓	19 2	81.9 0.9
Same as above	鬻, 鬻, 鬻, 鬻, 鬻, 鬻			
彌鬲	▢鬻〈叁〉 ▢▢弓〈叁〉弓鬲	鬻 ▥弓〈叁〉弓	19 3	74.6 1.9
彌鬲	▢鬻每 ▢▢弓每弓鬲	鬻 ▥弓每弓	19 3	74.6 1.9
彌鬲	▢鬻孝 ▢▢弓孝弓鬲	鬻 ▥弓孝弓	19 3	74.6 1.9
彌鬲	▢鬻咸 ▢▢弓咸弓鬲	鬻 ▥弓咸弓	10 1	82.6 0.9
彌鬲	▢鬻疢 ▢▢弓疢弓鬲	鬻 ▥弓疢弓	10 1	82.6 0.9
彌鬲	▢鬻采 ▢▢弓采弓鬲	鬻 ▥弓采弓	10 1	82.6 0.9

	⿱鬻去	鬻	10	82.6
鬻	⿲⿳弓去弓鬲	⿲弓去弓	1	0.9
	⿱鬻〈狀〉	鬻	10	69.4
鬻	⿲⿳弓〈狀〉弓鬲	⿲弓〈狀〉弓	2	2.8
	⿱鬻付	鬻	10	69.4
鬻	⿲⿳弓付弓鬲	⿲弓付弓	2	2.8

TABLE 11. ⿰⿱ A ⿲ LRC ↔ ⿱ A ⿲ LCR (611)

char	structure	component	pn	accuracy
〈斎〉	⿱〈齐〉貝	〈齐〉	30	93.7
	⿱文⿲丿貝丨	⿲丿貝丨	1	0.1
〈齑〉	⿱〈齐〉韭	〈齐〉	30	93.7
	⿱文⿲丿韭丨	⿲丿韭丨	1	0.1
Same as above	〈斎〉, 〈斎〉, 〈斎〉			
斎	⿱〈齐〉示	〈齐〉	30	82.6
	⿱文⿲丿示丨	⿲丿示丨	3	0.9
〈羸〉	⿱〈亯〉〈糸〉	〈亯〉	23	91.8
	⿱〈亯〉⿲〈月〉〈糸〉〈凡〉	⿲〈月〉〈糸〉〈凡〉	1	0.2
〈贏〉	⿱〈亯〉〈贝〉	〈亯〉	23	91.8
	⿱〈亯〉⿲〈月〉贝〈凡〉	⿲〈月〉贝〈凡〉	1	0.2
Same as above	贏, 〈贏〉, 〈贏〉, 〈贏〉, 〈贏〉, 〈贏〉, 〈贏〉, 〈贏〉, 〈贏〉, 〈贏〉, 〈贏〉, 〈贏〉, 〈贏〉, 〈贏〉			
〈羸〉	⿱〈亯〉馬	〈亯〉	23	84.6
	⿱〈亯〉⿲〈月〉馬〈凡〉	⿲〈月〉馬〈凡〉	2	0.7
羸	⿱亯魚	亯	16	88.6
	⿱亯⿲〈月〉魚凡	⿲〈月〉魚凡	1	0.4
羸	⿱亯叟	亯	16	88.6
	⿱亯⿲〈月〉叟凡	⿲〈月〉叟凡	1	0.4
Same as above	羸, 羸, 羸, 羸, 羸, 羸, 羸			
〈齎〉	⿱齊〈⿲月/冃/月/月〉	齊	7	76.6
	⿱㐬⿲丿〈⿲月/冃/月/月〉丨	⿲丿〈⿲月/冃/月/月〉丨	1	1.6
齎	⿱齊魚	齊	7	60.5
	⿱㐬⿲丿魚丨	⿲丿魚丨	2	5.0

〈齏〉	▢齊魚 ▢旂▢丿魚丨	齊 ▢▢丿魚丨	7 2	60.5 5.0
蠃	▢蠃虫 ▢亠▢〈月〉虫凡	蠃 ▢〈月〉虫凡	7 1	76.6 1.6
羸	▢蠃羊 ▢亠▢〈月〉羊凡	蠃 ▢〈月〉羊凡	7 1	76.6 1.6
蠃	▢蠃〈果〉 ▢亠▢〈月〉〈果〉凡	蠃 ▢〈月〉〈果〉凡	7 2	60.5 5.0

TABLE 12. ▢▢ AM ▢ LRC ↔ ▢ AM ▢ LCR (612)

char	structure	component	p*n*	accuracy
〈贏〉	▢〈蠃〉貝 ▢〈亡〉口▢〈月〉貝〈孖〉	〈蠃〉 ▢〈月〉貝〈孖〉	4 1	64.0 4.0
〈羸〉	▢〈蠃〉羊 ▢〈亡〉口▢〈月〉羊〈孖〉	〈蠃〉 ▢〈月〉羊〈孖〉	4 1	64.0 4.0
〈嬴〉	▢〈蠃〉女 ▢〈亡〉口▢〈月〉女〈孖〉	〈蠃〉 ▢〈月〉女〈孖〉	4 1	64.0 4.0

In many cases, especially in the trivial ones, the accuracy of the functional structures is high. Often these results are also consistent with the etymological structure. For example, in the case of "俻" (table 2), the accuracy of ▢攸田 is lower than the accuracy of ▢亻备. In fact, ▢亻备 seems to correspond to the etymological structure. However, since these results were calculated with respect to a simple count of the number of CHISE character objects, productivity of a component was divided by the number of glyph-variants of the component. For example, component 〈蠃〉 is written in various forms such as "蠃," "蠃," "蠃," "蠃," "蠃," "蠃," "蠃," "蠃," "羸," "蠃," etc. As the result, the productivity of apparent components (▢月*x*凡, ▢月*x*孖, etc.) increases and the accuracy of the functional structure decreases. Therefore, if information on variant-relations is available, it is better to normalize the glyphs (close to the original forms), its structure and components.

5. Conclusion

Structural description of Chinese character should be based on Chinese character analysis (Chinese character studies), like grammatical analysis of natural language. For this reason, it is desirable to describe Hanzi structure based on etymological explanations. Etymological knowledge

is required to perform the task, however this information is missing for many atypical Chinese characters, in fact the *majority* of CJKV Unified Ideographs of UCS.

On the other hand, if we consider the "(degree of) componentness" of Chinese characters, a (candidate for) component that produces more Chinese characters is more likely to be considered as an actual component. Considering components with respect to productivity does *not* require any etymological knowledge on characters and therefore can be calculated whenever we have a dataset of Hanzi structure descriptions at our disposal.

Based on this hypothesis, we conducted an experiment using the structural data on Chinese characters from the CHISE character ontology. As a result, we found that components based on etymological knowledge are more likely to have a higher productivity.

References

Asahara, Tatsuro [浅原達郎] (1996). 漢字の字符 *[Chinese Character Graphemes]*. URL: http://yuetgu.zinbun.kyoto-u.ac.jp:8098/yg/rs/jihu.pdf.

Information technology—Universal Coded Character Set (UCS) (2014). ISO/IEC 10646:2014. International Organization for Standardization (ISO).

Morioka, Tomohiko [守岡知彦] (2015). "Multiple-policy Character Annotation based on CHISE." In: *Journal of the Japanese Association for Digital Humanities* 1.1, pp. 86–106.

———— (2018a). "Integration of a Chinese Character Ontology Title and Historical Glyph Examples." In: *9th International Conference of Digital Archives and Digital Humanities (DADH 2018)*. Taiwanese Association for Digital Humanities / Dharma Drum Institute of Liberal Arts, pp. 287–300.

———— (2018b). "項書き換え系を用いた漢字字体の包摂規準の形式化の試み [An Attempt to Formalize Unification Rules of Chinese Characters Based on Term Rewriting System]." In: 情報処理学会論文誌 *[Journal of the Information Processing Society of Japan]* 59.2, pp. 332–340.

Slaměníková, Tereza (2019). "On the Nature of Unmotivated Components in Modern Chinese Characters." In: *Proceedings of Graphemics in the 21st Century, Brest 2018*. Ed. by Yannis Haralambous. Brest: Fluxus Editions, pp. 209–226.

The Development
of the Description of Punctuation
in Historical Grammar Books

Tomislav Stojanov

Abstract. The central thesis of this paper is that the evolution of punctuation reveals many interesting sociolinguistic aspects of a language. Studying punctuation can offer us new insights into the development of language codification, the relationship between speech and writing, the socio-cultural circumstances of a specific epoch, and it can even contribute to contemporary descriptions of orthographic prescription. On the history of the development of punctuation, several key titles have been written that take into consideration different text sources, periods and languages, e.g., Parkes (1992), Salmon (1999), and Mortara Garavelli (2008). In this paper I aim to contrastively explore descriptions of punctuation found exclusively in a selection of prototypical and accessible grammar books from the time of Antiquity to the Enlightenment through the perspective of historical 'comparative standardology' (Joseph, 1987), an approach that Deumert (2003, p. 1) claim has rarely been explored systematically. I have analysed five grammar books from Antiquity, fourty from Renaissance Humanism (twenty-one Latin and nineteen vernacular), and twelve grammar books from the Enlightenment. The three analysed factors—the grammar-book function, the divergence of punctuation from grammatical teaching to orthographic content, and the transformation of punctuation into written characters—were recognized as the most significant legacies of grammar books in the evolution of punctuation and in its transformation into the function, status, and application as we know it today. Supported by the constant evolution of literacy, the number of punctuation marks has been steadily increasing. The observation of historical punctuation in de jure and de facto normative orthographies or grammar books shows the strong link between the socio-cultural context and punctuation-related descriptions or prescriptions.

1. Introduction

Not all modern European languages have government-authorized orthographic manuals that standardize writing. Written standards have been

Tomislav Stojanov   ⓘD 0000-0002-6972-6518
1. Institute of Croatian Language and Linguistics, Zagreb, Croatia
2. University of Nottingham, School of Cultures, Languages and Area Studies, UK
E-mail: tstojan@gmail.com

established lexicographically either through general lexicography (e.g., English, French, Italian, and Spanish) or separate spelling dictionaries (e.g., Danish, German, Slovenian, and Russian). However, this was not the case throughout the history of language. Grammar books have assumed a central role in the process of language codification. Therefore, old grammar books provide numerous interesting insights into culture, the history of education, the evolution of linguistic thought, etc. (Law, 1997). By observing the history of punctuation in old grammar books through various socio-cultural circumstances and other factors in different epochs, we are able to learn more about the evolution of language description and prescription. A still-present practice of dividing punctuation marks into two classes—at sentential and word level—can be explained by an ancient differentiation between *distinctiones* and *notae*.[1] Furthermore, the principles on the basis of which contemporary punctuation norms have been established (cf. Salmon (1962, p. 348) and Salmon, 1988), have their origin in the rhetorical and grammatical function of punctuation.[2]

The status and meaning of orthography and punctuation have changed throughout history. Orthography used to be a constituent part of many historical grammar books, whereas descriptions of punctuation were more seldom. For example, the majority of the grammar books of Latin in Renaissance Humanism analysed here considered orthography a component of grammar, with its own unit named *littera*. In order to quantitatively depict the relations between orthography, punctuation, and other grammatical entities, I will here present the search results of several key words in the monumental *Lexicon Grammaticorum* that spans 1,728 pages. The word punctuation has 34 instances, orthography with its derivatives appears 485 times (spelling 270), as compared with 656 occurrences of morphology, 780 of semantics, 1,305 of phonetics, 1,236 of dictionary and 1,385 of syntax. This roughly exposes which topics have been dominantly linked with descriptions made by the world's most representative grammarians.

The most comprehensive exploration of the development of historical punctuation in Europe can be found in Parkes (1992) and Mortara Garavelli (2008). Wingo (1972) wrote an excellent treatise of Latin punctuation in the Classical Age. The evolution of English punctua-

---

1. One example is Babić, Finka, and Moguš (2004), a standard orthographic manual of Croatian. The sentential (Croatian *rečenični znakovi*) and the orthographic marks (Croatian *pravopisni znakovi*) are the same characters (e.g., period, comma, colon, etc.) with the difference that the first are separate sentences and the latter affect the pronunciation or meaning of a word.

2. For instance, one of the disputes during the 1960 Novi Sad spelling reform of Croatian was a switch in the prescription of the use of punctuation from a 'grammatical principle' to a logical ('free' or rhetorical) principle. Cf. Jonke (1962) and, earlier, Guberina (1940).

tion from 1476 to 1776 has been analysed by Salmon (1999).[3] Both latter sources included grammar books in their studies and points for further reading. However, a contrastive exploration of the description of punctuation specifically in grammar books, as the most influential representations of written norms in language history, is still lacking.

The most common way to categorize grammar books is in terms of the well-known historical epochs of Western civilization. From the punctuational point of view, this periodized approach yields an unwanted gap in the typology, since grammar books, such as those written under the influence of Rationalism (the so-called universal or the philosophical grammars), do not include descriptions of punctuation at all. Therefore, I have adopted a more specific classification, proposed by Vogl (2012: 22), which was created to depict the emergence of a standard language ideology:

1. the emergence of 'uniform written languages' in the Middle Ages;
2. the emergence of a 'correctness ideology' in Early Modern times ('language and norm');
3. the instrumentalization of 'correct languages' as vehicles of identity politics and the politics of democratization in the eighteenth and nineteenth century ('language and nation');
4. the devaluation of everything non-standard in the nineteenth and twentieth century ('the best variety').

For a discussion on punctuation in grammar books, the fourth period is not relevant, since punctuation eventually became separated from grammatical teachings and reached its orthographical status in the third phase. This standard-language-ideology timeline corresponds to the historical socio-cultural periods. The emergence of 'uniform written languages' is linked to Antiquity and the Middle Ages; the emergence of a 'correctness ideology' matches up with Renaissance Humanism; and the relationship between language and nation was established in the age of Enlightenment.

2.   Methodology

The selection criteria for grammar books was defined according to the grammar-book prototypicality, availability, edition, and the language status today.

---

3. I would like to point out here more useful sources for the study of (historical) punctuation. Houston (2013) is an interesting and a well-written popular scientific book on the evolution of numerous punctuation marks from Antiquity to the modern era. One important source for modern grapholinguistic studies on punctuation and its typology is Gallmann (1985).

Prototypicality refers to an attempt to include the relevant grammar books for a specific period, language, or country. While searching for them, I have used many sources of both historical and modern linguistic literature, such as Walch (1716) and Law (1997) for Latin, Marsden (1796), Howland Rowe (1974), and Horst (2016) for vernacular languages, and Kovachich (1786), Marsden (1796), Swiggers (2001), and Haßler and Neis (2009), for the Enlightenment period. I found some sources during my own search.

Availability can be exemplified by the case of Dévai Bíró Mátyás, who (Kamusella, 2009, p. 122) defines as the first grammarian of Hungarian in 1538. I could not find this grammar book, and instead used Sylvester (1539), a book one year younger.

To gain a methodologically consistent picture of the state of grammar-book descriptions of punctuation in vernacular languages, I used the criterion of the first printed grammar books in 19 available languages. The abovementioned Howland Rowe (1974) was particularly useful because of his comparative research into the first vernacular grammars of the sixteenth and seventeenth century for 63 languages. I have adopted his methodology of grammar-book determination and the list of grammars with two differences. Instead of the Hungarian grammar of Molnár (1610), I studied the earlier Sylvester (1539), and instead of the Portuguese grammar of Barros (1539), I considered the seven-year-older grammar of Oliveira (1532).

The last criterion was that I reviewed only languages that are officially used on a national level in Europe today, excluding minority and regional languages. The consulted grammar books are listed in the primary bibliography. There are five grammar books from Antiquity, 40 from the Renaissance Humanist era (21 Latin and 19 vernacular), and 12 from the Enlightenment period, which describe a total of eight languages.

The text sources of the grammars from the period of Antiquity were the Corpus Grammaticorum Latinorum webpage (sadly unavailable for some time now,[4]), the Greek Wikisource page[5] with Dionysus Thrax's Grammar, the Documenta Catholica Omnia website,[6] Davidson (1874), Copeland and Sluiter's selection of translated texts published in 2012, Barney et al. (2006), and the Google Books service.

---

4. `http://kaali.linguist.jussieu.fr/CGL`.

5. `https://goo.gl/oyQRVB`.

6. `http://www.documentacatholicaomnia.eu/`.

3.   Antiquity

On punctuation, Aristotle commented in the fourth century BCE that, 'It is a general rule that a written composition should be easy to read and therefore easy to deliver'. The reason why he included a discussion on punctuation in his book on rhetoric was obvious—according to Aristotle, one had to be skilled in rhetoric to punctuate a text.[7] Since proper punctuation was considered a skill of knowledgeable people, it logically appeared in the oldest preserved grammar of Greek, at the turn of the second to the first century BCE, Dionysus Thrax's Art of Grammar (*Τέχνη γραμματική*). It has 25 parts, two of which relate to the punctuation content: part (IV) on signs for clauses, and part (V) on the difference between the period and comma in terms of the criterion of time, i.e., the pause. The longest clause was the period (Greek *περίοδος*), which was marked by a high dot; a medium-long clause was the colon (Greek *κῶλον*), which was marked by an intermediate dot; and a short clause was marked by an underdot, or the comma (Greek *κόμμα*). The three basic punctuation marks represent syntactical and rhetorical units that indicate the manner of speaking, since texts in Antiquity were written in a continuous series of capital letters without blanks (Lat. *scriptura continua*).

This was adopted by the Roman grammarian Aelius Donatus, who wrote two Latin grammars (350 CE). *Ars maior* is an extended work and includes the chapter "De distinctionibus," which discusses the three positions of the separator character: high (Lat. *distinctio*), low (Lat. *subdistinctio*), and middle (Lat. *media distinctio*). Priscianus Caesariensis wrote *Institutiones grammaticae* around the year 520 CE and did not describe punctuation marks or other written characters.

Even though it is not a grammar book as such, but a medieval encyclopaedia, Isidore of Seville's *Etymologiae* from the sixth to seventh century was a highly influential book (or rather collection of books), among which the first one was dedicated to grammar. It is divided into 44 chapters, three of which relate, more or less, to what we would today associate with punctuation: *De posituris* (XX), *De notis sententiarum* (XXI), and *De notis vulgaribus* (XXII). *De posituris* is about punctuation, although Isidore, according to Aristophanes and other grammar-book precursors, continues to consider the *comma*, *colon*, and *periodos* to be parts of sentences, which led Barney et al. (ibid., p. 74) to translate these terms as *clause*, *phrase*, and *sentence*. *De notis sententiarum* deals with 26 sentence marks of 'critical reading' (asterisk, paragraph, quotation marks, etc.). *De notis vulgaribus* describes symbols that mark syllables and words.

---

7. 'To punctuate Heracleitus is no easy task, because we often cannot tell whether a particular word belongs to what precedes or what follows it.' Both Aristotle's citations are translated by W. Rhys Roberts and found in Barnes (1991, p. 114).

In *Ars grammatica* (ca. 798), Alcuin divides grammar into 26 types, among which there are punctuation marks (*positurae*), critical marks (*notae*), orthography, etc. Alcuin does not list them, but defines them— punctuation marks are, in Copeland and Sluiter's (2012) translation, 'points to distinguish meanings'. Critical marks symbolize 'certain marks, either to abbreviate marks, or to express meanings; or they are used for a variety of reasons, such as the obelus <÷> in Holy Scripture, or the asterisk <*>'.

4.   Renaissance Humanism

4.1.   Punctuation in Latin Grammar Books

In order to analyse the status of punctuation in Latin grammar books, I have looked at three aspects: (1) the definition of a grammar, (2) the content of the orthographic chapter, and (3) the description of punctuation. A total of 21 Latin grammar books, from the oldest one, Nebrija's in 1481, to Golius's in 1636, have been reviewed in more detail in Table 1.

The authors of the 12 Latin grammar books consider orthography a constituent part of grammar, equal with prosody, etymology (i.e., morphology and word formation) and syntax, with their respective units *littera* (letter or sound), *syllaba* (syllable), *dictio* (word), and *oratio* (clause). These are Nebrija, Cochlaeus, Curio, Melanchthon, Ramus, Valerius, Crusius, Alvares, Caucius, Frischlinus, Sanctius, and Golius. *Littera* is the 'sound which becomes separate by writing' (*vox, quae scribi potest individual*, Nebrija), while Scioppius goes even further by identifying *littera* as the basic unit of orthoepy, which was a synonym of orthography. The teaching on *littera*, the basic unit of orthography, is fundamentally about sounds. The grammatical content of *littera* in grammar books in this period is a division of letters and sounds, the difference between the letters K and Q, Z, and Y, discussions over the letter X, double letters, the arrangement of letters, diphthongs, and pronunciation of consonants with the sound *h*, etc. However, in spite of the *littera* definition and description, many grammarians define orthography as the art of writing correctly (*ars [recte] scribendi*, e.g., Nebrija, Curio, Crusius, Frischlinus, Golius), with prosody, etymology, and syntax being described as an art of speaking correctly (*ars [recte] loquendi*).[8]

The orthographic content in a wider sense (including the annexed chapter by Camerarius in Melanchthon's grammar book) encompasses the following 12 units in Latin grammar books:

8. An overview of the orthography and grammar definitions from the reviewed period can be found in Haßler and Neis (2009, pp. 1716–1730).

1. teachings about *littera*;
2. the division of *distinctiones* into a comma, a colon, and a period;
3. marks (*notae*): question mark, exclamation mark, round brackets, diaeresis, hyphen between words, hypodiastole, accent marks;
4. apostrophe;
5. capital and minuscule letters;
6. the abbreviation of writing;
7. the division of words into syllables;
8. spelling variants (e.g., *ad/at*, *obstitit/opstitit*);
9. deviations in writing or general spelling mistakes;
10. rhetorical figures (*de figuris orthographicis*) and deviation from usual writing: *adjectio*, *detractio*, *transmutatio,* and *immutatio*;
11. three theoretical perspectives: tradition (*autoritate*), etymology (*notatione*), and correctness (*proportione*);
12. an orthographic glossary with a list of Greek names that were transferred into Latin differently.

Based on the description, on the characters that are included, and on its location in grammar books, punctuation teaching can be divided into four categories. These can even be named as stages in the evolution of punctuation. Each grammar book belongs to a single category, except for Valerius, Frischlinus, and Camerarius, which share features from the third and fourth categories.

(a) grammar books without a description of punctuation;
(b) grammar books that inherited a description of punctuation from Antiquity with three basic characters—comma, colon, and period;
(c) grammar books with five basic punctuation characters—the three abovementioned marks plus the question mark and parenthesis;
(d) grammar books with innovative approaches to the description of punctuation;

Ten grammarians belong to the first category (Aventinus, Brassicanus, Lancilotus, Linacre, Scaliger, Ramus, Alvares, Caucius, Sanctius, and Scioppius).

The interest of Latin grammarians and prominent orthographers during the Renaissance Humanist period was intrinsically bound to Greek and its written history, and they thus inherited teaching from the Antique period. Some Latin grammarians consistently followed the 'traditionalist' grammatical teachings on punctuation (Clenardus, Melanchthon, and Sanctius) and they all belong to the second category. All three authors described punctuation as a syntactic phenomenon—the three basic characters (comma, colon, and period) were within or immediately followed the syntactic chapter.

Alsted and Golius belong to the third category, with five punctuation marks (comma, colon, period, plus question mark and parentheses), which were explained both syntactically (*partes periodi*) and respi-

ratorily (*notae respirationes*). Both of them grouped the period, colon, and comma into the respiratory characters, while the question mark and parentheses were sentence characters of sound change (*notae mutationis soni*), as defined by Alsted. There are more grammarians that we associate with this group—Manutius, Frischlinus, Valerius, Camerarius, and Curio. Unlike Alsted and Golius, their description of punctuation went beyond the solely rhetorical or syntactic. Instead of a description of speech finiteness or perfection, which was a typical grammatical aspect of punctuation in Antiquity, the punctuation content was no longer in the syntactic part, but (1) among grammatical foundations—at the end of the first book on grammar essentials (Valerius); (2) at the end of the book (Manutius), together with accents and meter; (3) as part of the orthography chapter (Curio, Frischlinus), or within the orthography annex of the grammar book (Camerarius).

The last, fourth category of punctuation among the Latin grammarians happened when the punctuation set was enlarged with other characters. These are characters that denote pronunciation—accents, diaereses, apostrophes, marks for long and short syllables, and hyphens (Valerius, Golius). Furthermore, these characters signal an even stronger influence of the written language, which would become more obvious in vernacular grammar books: capital letters (Frischlinus) and paragraph marks, obelisks, and asterisk signs (Camerarius).

The period, colon, comma, question mark, and parentheses were fundamental features in punctuation descriptions found in sixteenth century Latin grammar books. The exclamation mark appeared much later—first in Alsted (1610), and then in Golius (1636), even though an 'effect of admiration' is mentioned in Manutius (1507)—an author who considerably influenced today's punctuation standards in his famous work as an early printer and typographer.

Just three grammarians described punctuation within orthography. The first was Curio (1546), and next came Camerarius, the author of the orthography chapter that featured as an annex in Melanchthon's grammar book. (Melanchthon did not consider punctuation part of orthography, however.) This annex was printed eight years before Aldus Manutius's *Orthographiae ratio* (1561) and can be regarded as one of the oldest printed orthographic manuals of Latin. The third grammarian was Frischlinus (1586).

4.2.   Punctuation of the First Vernacular Grammars

Latin continued to be the language of science in the fifteenth and sixteenth century, and so it was the starting point for describing vernaculars. Most vernacular grammar books used Latin as their metalanguage (12 out of 19). The teaching of vernacular grammars was completely in-

herited from Latin grammar books. One reason why vernacular grammars rely on Latin grammars so strongly probably lies in Law's explanation: the more the description of a language was similar to Latin, the more successful the grammars were (Law, 2003, p. 234). This is why the first Nordic grammar books were even literally translated pursuant to Donatus's *Ars minor* (Hovdhaugen et al., 2000, p. 10).

However, the status of punctuation in vernacular grammar books reveals an interesting pattern related to grammar-book function. While Antique grammars were oriented towards the native speaker, the vernacular grammars placed the foreign language speaker at their centre.[9] One aspect of language learning and tutoring found in vernacular grammars is their including a key to understanding the function of punctuation in them. I have analysed 19 vernacular grammars in their first editions in relation to one of the most important socio-cultural factors of that time—religion. Table 2 shows the language and metalanguage of grammar books, the religious background and information on the inclusion of the description of punctuation.

For a grammar book such as this, whose author was among the ranks of the Catholic Church and was working towards the ultimate goal of supporting (re)evangelization and spreading the faith, punctuation was of secondary importance. Grammar books were aimed at missionaries and priests who needed to learn the vernacular, and who were starting from Latin. Since the Jesuits were in charge of this process, they decided to typify the Latin grammar (Alvares 1572) and to complement it with data from local languages. If Alvares's grammar had had any punctuation-related content, this would certainly have been transferred to the vernacular grammars that were modelled on it. It did not because the written language was not vital knowledge for the Catholic Counter-Reformation or Revival, which prioritized preaching, i.e., the spoken language. Four Catholic grammar books were analysed, among which three did not have any description of punctuation—Portuguese (Oliveira 1532), Croatian (Kašić 1604) and Irish (Maolmhuaidh 1677). One exception is Albertus (1573), albeit with the important detail that Albertus converted from the Protestant to the Catholic faith five years prior to the book being printed, which tentatively suggests it was written under the influence of Protestantism and different socio-cultural circumstances.

Likewise, the practical reason of learning a new language underpinned the secular grammars. The spoken language was once again more important to pilgrims, traders, diplomats, and other travellers.

---

9. Law (1997, p. xi). This is valid for the grammars that employ the Latin metalanguage. For the others, which were written in vernaculars, Vogl (2012, p. 20) explains that 'these grammars were not meant for foreign language learners, but for speakers of (a variety of) the languages to whom the authors of the grammars wanted to teach a "correct" version of their mother tongues.'

TABLE 1. Overview of Latin grammar books and their punctuation content

Author	Is there a chapter that includes orthography and/or punctuation?	Does it include punctuation? If so, what is the content?
Nebrija (1481)	7 pages on orthography (*De erotymatis orthographia*) in book 3 of 5. In the later edition (1515) he added a chapter *De punctis clausularum* on one page after the last, fifth book.	Comma, colon, period (*nota punctus*), parenthesis, and *nota interrogationis*.
Manutius (1507)	6 pages (*De posituris*) as the last book chapter, which also describes syllables, meter, and accents with many punctuation references	Period, colon, comma, question mark. Brackets were given as an example, but not directly named.
Cochlaeus (1514)	orthography without punctuation on 9 pages (*Folio LXXVI*).	no
Aventinus (1515)	no	no
Brassicanus (1518)	no	no
Lancilotus (1518)	no	no
Linacre (1532)	no	no
Scaliger (1540)	no	no
Curio (1546)	2 pages (*De orthographia*) in book 4 of 5.	Punctuation (*distinctiones*— period, comma, colon, question mark, brackets) is described in the chapter of orthography.
Clenardus (1551)	1 page (*Partes periodi*) in the second part of the book + 19 annexed pages (*De orthographia*) at the end of the book.	The description of sentence parts (*partes periodi*) as period, colon, and comma is after the description of syntax and before the part on accents and syllables. The orthography chapter was written by Johannes Vasaeus and it does not include punctuation.
Melanchthon (1553)	4 pages (*De periodis*) + 2 (*De distinctionibus*) + 16 annexed pages (*De orthographia*) at the end of the book, out of which 2.5 pages are dedicated to punctuation.	Description of sentence parts (*periodus, comma, colon*) is immediately after the syntax, and is followed by a chapter on *distinctiones* (*subdistinctio, media distinctio* and *distinctio vocalis/finalis*). The orthography chapter in this edition was written by Joachim Camerarius. It includes sentence marks (*de notis distinctionum*), which are: period, comma, colon, question mark, and brackets. In one paragraph, Camerarius also mentions compound marks (*multiplices notae*): *paragraphus, asteriscus*, and *obeliscus*.

Author (year)		
Ramus (1559)		no
Valerius (1560)	1 page (*Quaedam de notis*) at the end of the first book on the basics of grammar, before etymology.	*Comma, colon, periodus, interrogatio, parenthesis, apostrophus, mark for long and short syllable, hypodiastole,* and *diaresis,* and four accent marks.
Crusius (1563)	5 pages (*De orthographia*) at the beginning of the second part of the book + 8 pages (*De distinctionibus, et compositione orationis*) after the chapter on verbs and before the description of the calendar.	*Subdistinctio, media distinctio, finalis distinctio, interpositio* (round brackets), and *nota interrogationis.*
Alvares (1572)		no
Caucius (1581)		no
Frischlinus (1586)	8 pages (*De orthographia et prosodia*) at the beginning of the book, which includes the section *De notis* on punctuation (half of the page).	Orthography is determined in a twofold (*duplex*) fashion: basic (*simplex*) - sounds and letters, and formed (*figurata*)—marks (*de notes*) and figures (*de figuris orthographicis*). Punctuation is explained as separation marks (*notae distinctionum*), which are six: comma (or *virgula*), colon, period (*punctus finalis*), question mark (*nota interrogationis*), brackets (*nota parentheseos*) and capital letters (*litera majuscula*).
Sanctius (1587)		no
Alsted (1610)	2 pages on punctuation (*De orationis distinctione*) in the chapter on syntax, behind the part on exclamations and before syntactic figures.	no[a] Punctuation is divided into primary and secondary. Primary are respiratory marks (*nota respirationis*): *virgula, periodus,* and *duo puncti.* Secondary are marks of sound change (*nota mutationis soni*): *parenthesis, signum interrogationis,* and *signum exclamationis.*
Scioppius (1628)		no
Golius (1636)	4 pages (*De orthographia*) with which the book begins + 4 pages (*De ratione interpungendi*) as an appendix to the book on syntax (*Appendix ad syntaxin prior*).	*Comma, colon,* and *periodus* are punctuation marks based on the criterion of breathing (*respiratio*). Semicolon is included as part of the colon. Other punctuation marks are *interrogationis, parentheseos, exclamationis, diaereseos,* and *connexionis* (a hyphen between words, e.g., *ante-malorum*).

a. *Tropos, periodos, cola, commata* are mentioned once in the third book on syntax as figures of the verb and the sentence.

Table 2. Review of vernacular grammars

Work	Language	Metalanguage	Religious background	Descr. of punct.
Nebrija (1492)	Spanish	Spanish	Secular	No
Giovanni Francesco Fortunio (1516)[10]	Italian	Italian	Secular	No
Barclay (1521)	French	English	Secular	No
Oliveira (1532)	Portuguese	Portuguese	Catholic	No
Optát et al. (1533)	Czech	Czech	Protestant	Yes
Sylvester (1539)	Hungarian	Latin	Secular	No
Statorius (1568)	Polish	Latin	Protestant	No
Albertus (1573)	German	Latin	Prot. > Cath.	Yes
Spieghel (1584)	Dutch	Dutch	Secular	No
Bohorič (1584)	Slovenian	Latin	Protestant	Yes
Bullokar (1586)	English	English	Secular	No
Kašić (1604)	Croatian	Latin	Catholic	No
Portius (1638)	Greek	Latin	Secular	No
Petraeus (1649)	Finnish	Latin	Protestant	No
Jónsson (1651)	Icelandic	Latin	Secular	No
Pontoppidan (1668)	Danish	Latin	Protestant	Yes
Maolmhuaidh (1677)	Irish	Latin	Catholic	No
Tiällmann (1696)	Swedish	Swedish	Protestant	Yes
Ludolf (1696)	Russian	Latin	Secular	No

None of the nine secular grammar books explored here contained descriptions of punctuation. Unlike Catholicism, Protestantism relied heavily on printing and on spreading the written word. In the period from 1521 to 1545, 30.2% out of 5,651 printed books related to the reformation, and 17.6% to the Catholic doctrine. In the first half of the reviewed period, as much as 46% of all printed books related to reformation (Crofts, 1985, p. 373). These vernacular grammars attached more importance to punctuation because reading also became an important purpose for using the language. Most of the first vernacular grammars (four out of six), whose authors belonged to the Protestant priesthood, contain a description of punctuation to a smaller or greater extent.

Except for the five basic Latin punctuation marks, three more were included in this period: the hyphen, semicolon, and exclamation mark. The Czech grammar book introduced a hyphen at the end of a line, which illustrates a typographical influence on punctuation and the next step towards its separation from speech. The Danish grammar was the first to include the semicolon and exclamation mark (*signum admirationis*). The number of pages with a description of punctuation rose:

---

10. The first edition dates from 1516, however, I have used the edition from 1545.

while punctuation was listed on one to two pages in Latin grammars, the vernacular grammar book contained punctuation descriptions spanning from two and half to six pages (Czech—five pages, Slovenian—five pages, Danish—six pages, Swedish—two and half pages). All the grammar books included punctuation in the chapters on orthography. There was no notable correlation between a vernacular grammar's metalanguage and the description of punctuation.

5. The Enlightenment

One of the most obvious manifestations of the Enlightenment in European countries was the introduction of mass and compulsory primary education (Prussia 1763 and the Habsburg Monarchy 1774) and the establishment of national language academies (the Netherlands—1766, Russia—1783, Spain—1713, and Sweden—1783) or ministries of education (Poland—1773) with the goal of issuing normative grammars and establishing prescriptions concerning language use.

Among the first grammar books commissioned by language academies or other authorities with the goal of being normative and authoritative, the Russian (Lomonosov 1757), Polish (Kopczyński 1778), and Swedish (Sahlstedt 1769) grammar books included a description of punctuation. The half page on Russian punctuation encompasses the five basic marks, together with the semicolon, hyphen, and exclamation mark. Punctuation was called 'line characters' (Russian *строчные знаки*) and described in the second part of the book *О чтении и правописании россiйскомъ* ('On the reading and spelling of Russian'). Punctuation marks were named 'orthographic marks' in Polish (*znamiona pisarskie*) or *notae ortographicae*, with the Latin explanation in brackets, and described across two-and-a-half pages in the third part on grammar *O Znamionach* ('On marks'). They were the same as in Russian, while also including three footnote marks (1, a, *). The description of punctuation in Swedish is included in the last, sixteenth part of the grammar book (Swedish: *Om Skiljetecknen och andra uti skrifwande brukliga*, 'On punctuation and other writing habits'). It spans two-and-a-half pages and does not include the question mark among the five basic marks, but does include the semicolon, apostrophe, and diaeresis. Sahlstedt did not use the term 'orthography' in his grammar book.

The first normative grammar of Spanish, *Gramática de la lengua castellana* (1771), and Dutch, *Nederduitsche Spraakkunst* (1805), both commissioned by their respective national academies, do not include punctuation, only because the normative orthographic manuals had already been published (*Orthographía Española* for Spanish in 1741 and Siegenbeek for Dutch in 1804).

The Prussian government commissioned Johann Christoph Adelung to create a school grammar, which appeared in 1781 with a highly structured chapter on orthography, which included punctuation-related content. Adelung's description of orthography is in terms of a completely independent unit that he placed at the end of his grammar. It appeared in a separate publication entitled *Grundsätze der Deutschen Orthographie* one year later (1782). His fourand-halfpage-long subchapter on punctuation is divided into three categories: the first includes the question mark and the exclamation mark, the second the period, colon, semicolon, and the comma, and the third the quotation marks, the hyphen (*Theilungszeichen*) as <⸗> or <->, round and square brackets, the ellipsis (*das Zeichen einer abgebrochenen Rede*), the en-dash (*Gedankenstrich*) or <−>, and the apostrophe.

The school reformer under the rule of Maria Theresia, Johann Ignaz Felbiger, issued a German normative grammar in 1774, which did not include content pertaining to orthography or punctuation because it came out in the same year as a separate, also normative orthographic manual (Felbiger's *Anleitung zur deutschen Rechtschreibung: zum Gebrauche der deutschen Schulen in den kaiserlich-königlichen Staaten* in 1774). Felbiger's grammar served as a template grammar and orthography in all official languages of the Habsburg Monarchy (Hungarian, Croatian, Romanian, Slovakian, and others). It was first published in bilingual editions, and later as an adapted translation.

Unlike the above grammar books, all of which were normative language manuals in their societies, the following selection of grammar books in other countries were used as de facto language textbooks. They all include a description of punctuation marks. The most influential English grammar books in the period of the Enlightenment were Brightland and Gildon (1711) and Lowth (1762), with the latter said to be the 'embodiment of prescriptive grammar' (Tieken-Boon van Ostade, 2000, p. 881). Brightland and Gildon's grammar is divided into four parts—letters, syllables, words, and sentences. Punctuation, or *Stops and Pauses in Sentences* is described on its own, in the eleventh chapter on three pages, within the fourth part of the book that consists of three chapters (after the chapter on sentences that precedes, and before the chapter on prosody that succeeds it). The punctuation described is the comma, colon, semicolon, full stop or point, question mark, wonder or admiration mark, parenthesis, hyphen (at the end of a line), apostrophe, a caret mark that signifies an unintentionally omitted word in writing or printing, a stroke or a long line instead of word(s) deliberately left out, index point <☞>, obelisk mark as a footnote sign <‡>, section mark <§>, asterisk <*>, quotation marks <" ">, and paragraph mark <¶>.

Lowth's chapter on punctuation, which is 17 pages long, is structurally equal to the other parts of the grammar book and is positioned at the end of the book. It includes the comma, colon, semicolon, period,

question mark, exclamation mark, and the parenthesis, without mentioning orthography. According to Stammerjohann (2009, p. 932), Ash (1763) was used in schools as an adaptation of Lowth's grammar. This grammar has a 3-page separate chapter on punctuation ('Of the Points and Stops, and Other Characters Made Use of in Writing') at the end of the introductory chapter entitled 'An Introduction to the Grammatical Institutes'. The term 'orthography' was not used. The punctuation marks included are the comma, semicolon, colon, period, question mark, exclamation mark, quotation marks (<' '> or <" ">), brackets, caret, hyphen, apostrophe, paragraph mark (¶), diaeresis, and marks for notes at the bottom of the page (<*>, <†>, <‡>, or <||>). Capital and minuscule letters are also mentioned here.

In America, Webster (1783−1785) wrote a grammar in three volumes: the first was dedicated to orthography (*Spelling Book*, 1783), the second to grammar (*Grammar*, 1784), while the third part was a reader (*Reader*, 1785). Punctuation-related content was included in two places: a one-page description, taken over from Brightland and Gildon's first book, with one slight change—the omission of the long line. The other description is in the appendix of the second book and spans six pages, with the subtitle 'Abridged from Dr. Lowth'. It includes the comma, semicolon, colon, period, question mark, exclamation mark, and the parenthesis.

Italy was not politically united in the eighteenth century, so no wide-ranging educational reforms for learning Italian could be completed. Corticelli (1745) was the first Italian grammar with a clear educational function. Punctuation is described in several subchapters in the last part of the book (In Italian: *Della maniera di pronunziare, e di scriver toscano*, 'On How to Pronounce and Write Tuscan' [i.e., Italian]). This third part was entitled *Della ortografia toscana* ('On Tuscan Orthography') in the page heading. Writing apostrophes was included in the fourth part, and writing periods and commas was in the eleventh chapter, which spanned a total of five pages. Besides the apostrophe, period, and the comma, only the question mark, exclamation mark, and the semicolon were described.

Based on the 12 reviewed de jure and de facto normative grammars in eight language environments (German and English in two political systems), nine of them describe punctuation marks (three English grammar books in England and one in America, German in Prussia, Polish, Russian, Swedish, and Italian), while three do not (Dutch, German in Austria, and Spanish). The reason why punctuation is not found in normative grammars in the Netherlands, the Habsburg Monarchy, and in Spain is that the orthographic content had already been separated from the grammatical teaching and had grown independently into a separate publication. The normative orthographic manuals were published alongside the normative grammars. Out of nine grammar books that included punctuation, five of them included it in the orthographic chap-

ter (Italian, English in America, German in Prussia, Russian) or, indeed, named punctuation marks 'orthographic marks' (Polish). The remaining four grammar books described punctuation in their own chapters, two of which made the punctuation chapter equal to other book parts or chapters (Sahlstedt and Lowth), whereas two grammar books categorized punctuation within the Introduction part (Ash) or together with the chapter on sentences and prosody. None of these four grammar books linked punctuation with orthography.

The Enlightenment grammar books introduced three major novelties in punctuation. First, the punctuation-related content has been created with pedagogical criteria in mind, so that the rules became more structured, shorter, and clearer. Second, punctuation has eventually become separate from speech. The written perspective taken to punctuation is visible in the inclusion of footnote marks, hyphens at the end of lines, square brackets, dashes, various quotation marks, etc. Third, punctuation has become an essential part of language prescriptions due to the orthographic content finally being separated from grammatical teachings.

6.   Conclusions

Grammar books, central manuals in the history of language description, were the first framework in which content related to punctuation was described. The description of punctuation has a long history in grammar books from Antiquity to the Enlightenment. As grammar books evolved in different epochs, the teachings included on punctuation also changed—this signifies that punctuation relates to the socio-cultural context of grammar books. In this comparative analysis of the description of punctuation in historical grammar books, I have shown that the development of punctuation can be divided into three historical periods, which generally correspond to the classification of the emergence of a standard language ideology (Vogl 2012). I have isolated three major factors in the evolution of punctuation: the grammar book function, the divergence of punctuation from grammatical teaching into orthographic content, and the transformation of punctuation into written characters.

6.1.   The Grammar-Book Function

The first factor is the change in the relationship between punctuation and grammar-book functions. The evolution of punctuation can be evaluated as the history of the change in function of the grammar book. Punctuation arose from a pragmatic purpose of consuming written texts. The aim of punctuation in the Classical Age was to show the

sentence structure in order to ease the clarity of the written text and to facilitate reading. For this purpose, three basic characters were enough.

The main function of grammar books in Renaissance Humanism was to help educate pupils in Latin, a language void of native speakers for centuries. This is why grammatical teaching was inherited from the period of Antiquity, when grammarians were native in Latin. Moreover, Renaissance Humanism was affected greatly by the ancient texts that came to Europe via trade routes with the East. All humanists were consumers of manuscripts and there is no humanism without books (Davies, 2004, p. 47). Some even say that Renaissance Humanism 'may be regarded as a primarily language-oriented (or "lingual") movement' (Verburg, 1998, p. 189). The first printed grammar books of Latin, Nebrija, and Manutius started to include other characters among the punctuation marks from Antiquity, namely, the question mark and parentheses. The turning point was in the middle of the sixteenth century with Melanchthon (1553) and Valerius (1560), after which no one considered punctuation marks to be only the period, the comma, and the colon.

Regarding the content of punctuation, the discovery of the printing press affected punctuation considerably and represented the next stage in its evolution. Printed texts were more dominant, and punctuation evolved into standardized typographical marks. The number of standard punctuation marks raised from three to at least five. Two new punctuation marks were introduced—brackets and the question mark.

Based on its own description, punctuation in Latin grammar books was categorized into four groups (cf. 4.1). Among those authors who include descriptions of punctuation, we can conclude that punctuation evolved when it had begun to be considered as speech-related marks, outside of the scope of syntax.

On the other hand, the growing importance of the vernacular languages in administration and literary activity led to the emergence of vernacular grammar books (Percival, 2007). A need to spread religion and to learn vernacular languages were the factors that explain the (mis)appearance of punctuation's description in the first vernacular grammar books in Renaissance Humanism. Only grammar books written under the influence of Protestantism included descriptions of punctuation, which reveals the written character of language and the purpose of the grammar books.

The Enlightenment brought with it the last phase in punctuation's evolution. The function of grammar books changed substantially: they became prescriptive manuals commissioned by language institutions. Descriptions of punctuation were included in all the researched grammar books with the above-explained exception of two grammar books in which punctuation-related content was already printed separately in an associated orthographic textbook. The most representative feature of the Enlightenment was the introduction of the system of compulsory

public education. New grammar books had to satisfy the need for mass literacy in writing and reading. This led to the inclusion of punctuation because the unavoidable written characters and the introduction of new punctuation marks emphasized the even stronger influence of the written language.

6.2. Punctuation's Shift From Grammatical Teaching to Orthographic Content

In Antiquity punctuation was included in grammar books because of its rhetorical-syntactic role and the need to delimit speech. *Positurae*, *distinctiones* or *théseis* (period, colon, and comma, or *subdistinctio*, *media distinctio,* and *distinctio finalis*) were syntactic units that represented different parts of the sentence in order to indicate a level of finiteness of expression. They were also rhetorical marks because they symbolized places to breathe in while reading the *scriptura continua* texts. This teaching was inherited by the Latin grammarians Clenardus, Melanchthon, Sanctius, Alsted, and Golius. Other Latin grammarians, such as Manutius and Valerius, described punctuation as speech characters outside the syntactic chapters, but nevertheless punctuation was part of grammatical teaching. The change in the conception of punctuation happened in the middle of the sixteenth century with three grammarians—Curio, Camerarius, and Frischlinus—who began to regard punctuation as related to orthography.

The link between punctuation and orthography is clearly visible among vernacular grammarians. All of the four grammarians who were influenced by Protestantism, included their description of punctuation within the chapter on orthography, unlike the Catholic and secular grammarians (with just one debatable exception). These grammarians enlarged the standard set of punctuation marks to include the hyphen at the end of a line, the semicolon, and the exclamation mark.

The final stage in the evolution of punctuation was the ultimate separation from grammatical teaching that happened during the Enlightenment. In two-thirds of the languages analysed here, punctuation was considered as part of orthography, described either within grammar books (Italian, English in America, German in Prussia, Russian, and Polish) or even completely separately in prescriptive orthographic textbooks (Dutch, German in the Habsburg Monarchy, and Spanish). Punctuation was mostly described as separate from other grammatical features in the remaining four grammar books too, but without any mention of orthography. This divergence from grammatical teachings in the Enlightenment was followed by the introduction of many new punctuation marks, which led to the next, final element in the evolution of punctuation.

The first mention of punctuation as orthographic marks was in the first Polish grammar book, Kopczyński (1778) (*notae ortographicae* or *znamiona pisarskie*), followed by Adelung (1781), who described punctuation (*interpunction*) as orthographic marks (*orthographische Zeichen*).

6.3.   The Transformation of Punctuation Into Written Characters

The modern classification categorizes punctuation separately from other written characters (cf. Gallmann, 1985, and the Unicode standard), such as letters, symbols, numbers, etc. The distinction between the spoken and the written language is one of the most important in the evolution of language theory. From the three basic punctuation marks in Antiquity, today we count 798 characters that fall under the 'General Category of Punctuation' in the Unicode standard.[11] It is not incorrect to say that this great progress in the number of punctuation characters was caused by the demands of contemporary literacy and frequent language use in the written form.[12]

In the periods considered, from Antiquity to Renaissance Humanism the spoken language was at the centre of grammatical descriptions, as can be seen in the definitions of grammar and the status of orthography and punctuation in it. The more a language was used in the written form, the more punctuation marks appeared in grammar books. This process was followed by the separation of punctuation from grammatical teaching, as explained in the previous section.

As with the two abovementioned described factors, there are three observable periods in the evolution of punctuation. A shift from handwriting to printed grammar books (or Antiquity to Renaissance Humanism) affected typographical standardization and the number of punctuation marks. Following the early printers, grammarians such as Camerarius and Frischlinus began to include new characters. They inserted capital letters and three text marks (paragraph mark, obelisk, and asterisk mark), and Optát et al. added a hyphen at the end of a line, while Golius, brought a hyphen inside a line to denote the structure of a word. All these new characters represent the increasing influence of the written language and printed books.

The second change that happened in the Enlightenment era was fostered by mass education and literacy. Being able to read and write became a requirement that led to the spread of the written language throughout many societal circles and with many applications. New

---

11. The Unicode Standard v13, `https://www.unicode.org/charts/`. Accessed on 8 September 2020.

12. Parkes (1992, p. 2) stated that punctuation developed by stages that coincided with changing patterns of literacy.

characters that appeared in prescriptive grammar books represented the influence of the written language—footnote signs, various dashes and quotation marks, square brackets, etc.

By observing the descriptions of punctuation in a selection of prototypical and accessible grammar books in different periods from Antiquity to the Enlightenment, and taking into account the function of the grammar book, the shift in punctuation from grammatical teaching to orthographic content, and the transformation of punctuation into written characters, a typology of the development of punctuation across three periods can be established:

1. The handwriting punctuation of Antiquity with three basic characters (period, comma, colon) that served a rhetorical-syntactic function and which were described within syntactic chapters.
2. The standardized punctuation of Renaissance Humanism with five basic characters (period, comma, colon, question mark, brackets) that served the roles of learning Latin and spreading the influence of vernacular languages in printed books. Punctuation reflected both the spoken and the written language, and it was described predominantly outside the syntactic chapters.
3. The prescribed punctuation of the Enlightenment with more than eight punctuation marks that served the role of learning a national language as part of mandatory education and with the aim of increasing literacy. The punctuation reflects the written language and is included as part of orthographic content.

This paper aims to contribute to the description and typology of punctuation (based on Vogl's classification of a standard language ideology) and to the recognition of comparative (historical) standardology, as defined by Joseph. I have shown that punctuation went through three major evolutionary periods that evidenced the emergence of uniform written languages, the emergence of normative written languages, and the establishment of prescriptive written languages.

The three analysed factors or in other terms—punctuation function (6.1), punctuation status (6.2) and punctuation application (6.3)—can be recognized as the most significant legacies of grammar books in the history of punctuation in relation to punctuation's transformation into the forms and meaning with which we are familiar today. The function is depicted by the change from Latin to vernacular languages, the status by the inclusion of punctuation in orthographic content, and the application by the use of characters that represented the printed language. A pioneering grammar book is finally worth mentioning here—the first Czech grammar (Optát et al. 1533), which apart from being the first vernacular grammar that included the description of punctuation (punctuation function), it also included it within orthography (punctuation status), and added the hyphen at the end of a line as a character of printed language (punctuation application).

Acknowledgements

Primary Sources

Antiquity

Alcuin. *Ars grammatica* (∼ 790 AD).
Dionysius Thrax [Διονύσιος Θράξ]. *Τέχνη γραμματική* [*Art of Grammar*] (2nd c. BC).
Aelius Donatus. *Ars minor—de partibus orationis* (∼ 360–380 AD).
Aelius Donatus. *Ars maior* (∼ 360–380 AD).
Isidore of Seville. *Etymologiae* (∼ 600–625 AD).
Priscianus Caesariensis. *Institutiones grammaticae* (∼ 500–530 AD).
Quintilian. *Instituio Oratoria* (∼ 95 AD).

Renaissance Humanism

Albertus, Laurentius (Augsburg 1573). *Teutsch Grammatick oder Sprachkunst*.
Alsted, Johann Heinrich (Herborn 1610). *Compendium grammaticae latinae*.
Alvares, Emmanuel (Lisbon 1572). *De institutione grammatica libri tres*.
Aventinus, Johannes (Augsburg 1515). *Grammatica nova fundamentalis juvenibus utilissima*.
Barclay, Alexander (London 1521). *Here begynneth the introductory to wryte, and to pronounce French*.
Barros, João de (Lisbon 1539). *Grammatica da lingua portuguesa*.
Bohorič, Adam. (Wittenberg 1584). *Articæ horulæ succissivæ de Latinocarniolana literatura*.
Brassicanus, Joanis (Leipzig 1518). *Institutiones grammaticae elimatissimae*.
Bullokar, William (London 1586). *Bref grammar for English*.

Caucius, Antonius (Leiden 1581). *Grammatica latina*.

Clenardus, Nicolaus (Braga 1551). *Institutiones grammaticae Latinae*.

Cochlaeus, Johannes (Nuremberg 1514). *Grammatica rudimenta ad latinae linguae*.

Crusius, Martinus (Basel 1563). *Grammatica graeca, cum latina congruens*.

Curio, Coelius Secundus (Basel 1546). *De literis doctrinaque puerili, libri quinque*.

Fortunio, Giovanni Francesco (Venice 1516). *Regole grammaticali della volgar lingua*.

Frischlinus, Nicodemus (Frankfurt a.M. 1586). *Grammatica latina*.

Golius, Theophil (Strasbourg 1636). *Grammatica latina*.

Jónsson, Runólfur (Copenhagen 1651). *Recentissima antiquissimae linguae septentrionalis incunabula; id est grammaticae Islandicae rudimenta*.

Kašić, Bartol (Rome 1604). *Institutionum linguae Illyricae libri duo*.

Lancilotus, Curius (Strasbourg 1518). *De arte grammatica libri octo*.

Linacre, Thomas (Venice 1532). *De emendata structura latini sermonis libri VI*.

Ludolf, Heinrich Wilhelm (Oxford 1696). *Grammatica russica*.

Maolmhuaidh, Froinsias Ó [O'Molloy, Francis] (Rome 1677). *Grammatica Latino-hibernica nunc compendiata*.

Melanchthon, Pilip (Nuremberg 1553). *Grammatica latina*.

Molnár, Albert Szenczi (Hanau 1610). *Novae grammaticae Ungaricae*.

Nebrija, Elio Antonio de (Salamanca 1481). *Introductiones latinae cum commento*.

Nebrija, Elio Antonio de (Salamanca 1492). *La gramática que nuevamente hizo el maestro Antonio de Lebrixa sobre la lengua castellana*.

Nebrija, Elio Antonio de (Salamanca 1515). *Grammatica cu[m] quarta editione*.

Manutius, Aldus (Venice 1507). *Institutionum grammaticarum libri quator*.

Oliveira, Fernão de (Lisbon 1532). *Grammatica da lingoagem portuguesa*.

Optát, Václav Beneš, Gzel, Petar, Philomathes, Václav (Náměšť nad Oslavou 1533). *Grammatyka česká v dvojí stránce*.

Petraeus, Eskil (Turku 1649). *Linguae Finnicae brevis institutio*.

Pontoppidan, Erik Eriksen (Copenhagen 1668). *Grammatica danica*.

Portius, Simon (Paris 1638). *Γραμματικὴ τῆς Ῥωμαϊκῆς Γλώσσας. Grammatica linguæ Græcæ vulgaris*.

Ramus, Petrus (Paris 1559). *Scholae gramaticae*.

Sanctius, Franciscus Brocensis (Salamanca 1587). *Minerva sive de causis linguae latinae*.

Scaliger, Julius (Lyon 1540). *De causis linguae latinae libri tredecim*.

Scioppius, Gaspar (Franeker 1628). *Grammatica philosophica*.

Spieghel, Hendrick Laurenszoon (Asmterdam 1584). *Twe-spraack vande Nederduitsche letterkunst*.

Statorius, Pierre (Köln 1568). *Polonicae grammatices institutio*.

Sylvester, János (Sárvár 1539). *Grammatica hungarolatina*.

Tiällmann, Nils (Stockholm 1696). *Grammatica Suecana åller eñ Svensk Sprak-ock Skrif-konst.*

Valerius, Cornelius (Köln 1560). *Grammaticarum institutionum libri IIII.*

The Enlightenment

Adelung, Johann Christoph (Berlin 1781). *Deutsche Sprachlehre. Zum Gebrauche der Schulen in den Königl. Preuss. Landen.*

Ash, John (Worcester 1763). *Grammatical Institutes.*

Brightland, John, Gildon, Charles (London 1711). *A Grammar of the English Tongue.*

Corticelli, Salvatore (Bologna 1745). *Regole ed osservazioni della lingua toscana.*

Felbiger, Ignjat (Vienna/Freyburg 1774). *Anleitung zur deutschen Sprachlehre. Zum Gebrauche der deutschen Schulen in den kaiserlich-königlichen Staaten.*

Kopczyński, Onufry (Warsaw 1778). *Grammatyka języka polskiego i łacińskiego dla szkół narodowych na klasę pierwszą,* Warsaw, 1780 (*… na klasę drugą*); 1781 (*… na klase trzecią*).

Lomonosov, Mikhail Vasilyevich [Ломоносов, Михайл В.] (Sankt-Petersburg 1757). *Россійская грамматика* [*Russian grammar*].

Lowth, Robert (London 1762). *Short Introduction to English Grammar.*

Sahlstedt, Abraham (Upsala 1769). *Swensk Grammatika.*

Webster, Noah (Hartford 1783). *A Grammatical Institute of the English Language. Spelling Book.* 1784. *Grammar* (1785). *Reader.*

Weiland, Pieter (Amsterdam 1805). *Nederduitsche Spraakkunst.*

References

Babić, Stjepan, Božidar Finka, and Milan Moguš (2004). *Hrvatski pravopis [Croatian Orthography].* 8th ed. Zagreb: Školska knjiga.

Barnes, Jonathan, ed. (1991). *The Complete Works of Aristotle. The Revised Oxford Translation.* Vol. 2. Princeton, NJ: Princeton University Press.

Barney, Stephan A. et al. (2006). *The Etymologies of Isidore of Seville.* Cambridge, UK: Cambridge University Press.

Copeland, Rita and Ineke Sluiter (2012). *Medieval Grammar and Rhetoric: Language Arts and Literary Theory, AD 300–1475.* Oxford, UK: Oxford University Press.

Crofts, Richard A. (1985). "Printing, Reform, and the Catholic Reformation in Germany (1521–1545)." In: *Sixteenth Century Journal* 16.3, pp. 369–381.

Davidson, Thomas (1874). *The Grammar of Dionysios Thrax. Translated from the Greek.* St. Louis, MO.

Davies, Martin (2004). "Humanism in script and print in the fifteenth century." In: *The Cambridge Companion to Renaissance Humanism*. Ed. by Jill Kraye. Cambridge, UK: Cambridge University Press, pp. 47–62.

Deumert Ana; Vandenbussche, Wim (2003). "Standard languages: Taxonomies and histories." In: *Germanic Standardizations. Past to Present*. Amsterdam: John Benjamins, pp. 1–14.

Gallmann, Peter (1985). *Graphische Elemente der geschriebenen Sprache*. Tübingen: Niemeyer.

Guberina, Petar (1940). "Interpunkcija "novoga" (Belićeva) pravopisa u svijetlu logike i stilistike [Punctuation of the 'new' (Belić) orthographic manual in the light of logic and stylistics]." In: *Nastavni Vjesnik* 48.1939–1940, pp. 329–351.

Haßler, Gerda and Cordula Neis (2009). *Lexicon sprachtheoretischer Grundbegriffe des 17. und 18. Jahrhunderts*. Berlin: de Gruyter.

Horst, Joop van der (n.d.). *Propast standardnoga jezika. Mijena u jezičnoj kulturi Zapadne Europe [The Destruction of the Standard Language. A shift in the language culture of the Western Europe]*. Translated into Croatian by Radovan Lučić. Original title in Dutch: *Het einde van de standaardtaal. Een wisseling van Europese taalcultuur*. Zagreb: srednja europa.

Houston, Keith (2013). *Shady Characters. The Secret Life of Punctuation, Symbols & Other Typographical Marks*. New York: W. W. Norton & Company.

Hovdhaugen, Even et al., eds. (2000). *The History of Linguistics in the Nordic Countries*. Helsinki: Societas Scientiarum Fennica.

Howland Rowe, John (1974). "Sixteenth and Seventeenth Century Grammars." In: *Studies in the History of Linguistics. Traditions and Paradigms*. Ed. by Dell Hymes. Bloomington, IN: Indiana University Press, pp. 361–379.

Jonke, Ljudevit (1962). "Načela i primjena logičke interpunkcije [Principles and application of logical punctuation]." In: *Jezik* 9.3, pp. 71–78.

Joseph, John E. (1987). *Eloquence and Power. The rise of language standards and standard languages*. London: Frances Pinter.

Kamusella, Tomasz (2009). *The Politics of Language and Nationalism in Modern Central Europe*. New York: Palgrave Macmillan.

Kovachich, Martin Georg (1786). *Merkur von Ungarn: oder Litterarzeitung für d. Königreich Ungarn u. dessen Kronländer*. Pest: Gesellschaft patriotischer Liebhaber der Litteratur.

Law, Vivien (1997). *Grammar and Grammarians in the Early Middle Ages*. London: Longman.

———— (2003). *The History of Linguistics in Europe. From Plato to 1600*. Cambridge, UK: Cambridge University Press.

Marsden, William (1796). *A Catalogue of Dictionaries, vocabularies, grammars, and Alphabets. In Two Parts*. London.

Mortara Garavelli, Bice, ed. (2008). *Storia della punteggiatura in Europa*. Bari: Editori Laterza.

Parkes, Malcolm Beckwith (1992). *Pause and Effect. An Introduction to the History of Punctuation in the West*. Farnham, Surrey: Ashgate.

Percival, Keith W. (2007). "Grammar, Humanism, and Renaissance Italy." In: *Mediterranean Studies* 16.2007, pp. 94–119.

Salmon, Vivian (1962). "Early Seventeenth-Century Punctuation as a Guide to Sentence Structure." In: *The Review of English Studies* 13.52, pp. 347–360.

——— (1988). "English Punctuation Theory 1500–1800." In: *Anglia—Zeitschrift für englische Philologie* 106.3–4, pp. 285–314.

——— (1999). "Orthography and Punctuation." In: *The Cambridge History of the English Language*. Ed. by Roger Lass. Vol. Volume III. 1476–1776. Cambridge, UK: Cambridge University Press, pp. 13–55.

Stammerjohann, Harro, ed. (2009). *Lexicon grammaticorum. A Bio-Bibliographical Companion to the History of Linguistics*. 2nd ed. Tübingen: Max Niemeyer.

Swiggers, Pierre (2001). "Grammar." In: *Encyclopedia of the Enlightenment*. Ed. by Michel Delon. Vol. I A-L. London: Routledge, pp. 617–622.

Tieken-Boon van Ostade, Ingrid (2000). "Normative studies in England." In: *History of Language Science. An International Handbook on the Evolution of the Study of Language from the Beginnings to the Present*. Ed. by Sylvain Auroux et al. Vol. Volume 1. Berlin: de Gruyter, pp. 876–887.

Verburg, Pieter Adrianus (1998). *Language and its Functions*. Amsterdam: John Benjamins.

Vogl, Ulrike (2012). "Multilingualism in a standard language culture." In: *Standard Languages and Multilingualism in European History*. Ed. by Matthias Hüning et al. Amsterdam: John Benjamins, pp. 1–42.

Walch, Johann Georg (1716). *Historia critica latinae linguae*. Lipsiae: sumtu I. F. Gleditschii B. filii.

Wingo, Otha E. (1972). *Latin Punctuation in the Classical Age*. The Hague: Mouton.

Form-Meaning Regularities
in Old English Lexicon

Nataliia Drozhashchikh · Elena Efimova · Evgenia Meshcheryakova

Abstract. This article deals with the form-meaning hypothesis in Old English within the theory of arbitrariness/non-arbitrariness. It focuses on the relations between the initial grapheme (phoneme) in a word and its lexical semantics and aims to reveal any non-arbitrary form-meaning associations at the lexical level. The data include <w>-, <s>-, <h>-, and <p>-words from a Thesaurus of Old English. The methodology employs statistical methods (Chi-square test, the coefficient of contingency, the contributions to the Chi-square) within Python realization. Our primary hypothesis is that alliteration—regular repetition of onsets in Old English lexemes, could stand for the regularities in the semantics of these words. We extrapolate the initial research and underlying hypothesis to lexical data in general. The findings demonstrate non-arbitrary form-meaning regularities at the level of the entire Old English lexicon—the tendency of words sharing initial graphemes to be attracted to certain semantic categories.

1. Introduction

The present study focuses on form-meaning relationships in the Old English lexicon. Form-meaning mappings have long been in the focus of attention in semiotics and linguistics. They were first mentioned in the theory of the correctness of names in Plato's *Cratylus*. According to *phusei* approach, it was proposed that "there is a kind of inherent correctness in names" and according to *sunthēkē and homologia* approach, it

Nataliia Drozhashchikh   0000-0002-5910-2402
Tyumen State University, Volodarskogo 6, 625003 Tyumen, Russia
E-mail: n.v.drozhashhikh@utmn.ru

Elena Efimova   0000-0002-6584-965X
Tyumen State University, Volodarskogo 6, 625003 Tyumen, Russia
E-mail: elena.efimova@gmail.com

Evgenia Meshcheryakova   0000-0001-8748-640X
Independent researcher
E-mail: evg.meshch@gmail.com

was believed that there is no "correctness of names other than convention and agreement".[1] The phusei approach has served as the basis for developing the theory of non-arbitrariness (iconicity) that includes into its scope all motivated form-meaning cases (onomatopoeia, sound symbolism, phonaesthemes, ideophones, etc.). In recent decades an increasing number of publications on non-arbitrary coding in language have seemed to undermine the thesis of linguistic conventionality and have made it clear that there cannot be a place for "convention and agreement" dogma in linguistic theory. While studying frequency and complexity of letters and script (Altmann, 2004); universal lexical semantics across languages (Blasi et al., 2016; Youn et al., 2016); phonological systematicity (Monaghan, Shillcock, Christiansen, and Kirby, 2014); correlation of orthographic/phonological form (Jee, Tamariz, and Shillcock, 2018); form-meaning mapping in alliterative verse (Cornell, 1981); semantic functions of phonemic clusters (Lvova, 2005); the semantics of graphemes (Slaměníková, 2019), modern scholars provide evidence that language combines both arbitrary and non-arbitrary relations.

Form-meaning relations in earlier stages of language have not been sufficiently studied. A few papers focus directly on this kind of relationship in Old English, in particular, on sound symbolism, submorphemic iconicity, and form-meaning association in alliteration (Cornell, 1981; Jespersen, 1922; Minkova, 2003; Philps, 2008; C. A. Smith, 2016). While studying the relations between the phonemes and the semantic classes like "sound," "tone," "size," "movement," "human body," etc., the authors ascribe the meaning to the phonemic clusters themselves. E.g., Philps (2008) claims that word-initial phonaesthemes "are endowed with a potential for meaning". Cornell (1981) does not attribute meaning to the phonemes but reveals the tendency of alliterating sounds to be connected with certain connotative meanings. As far as the entire Old English lexicon is concerned, the research pertaining to the association of the initial phoneme in the words and their referential meaning is insufficient. Meanwhile, referentiality is directly related to the process of non-arbitrary coding in older languages as early nominations are more iconic (see Atkinson, Mills, and Smith, 2019).

Historical linguistics and the theory of evolution put an important emphasis on the role of non-arbitrary coding in language formation and development. Older languages represent the phase of development, where a clear motivation of the nomination and word formation processes are possible, and the original etymological basis of referential meanings has not yet been suppressed. The formation processes that take place in different parts of the language system at the earlier stages of its development provide more information about possible form-meaning relationships than in modern languages. For example, it

---

1. http://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.01.0172

is known that as the vocabulary grows more abstract, the original meaning of the word, which, as a rule, is concrete, becomes obscure. The growing number of derivatives makes it difficult to extract monomorphic lexemes that retain their original non-derivational meanings. The penetration and assimilation of lexical borrowings in the receiving language complicate the distinction between the original and borrowed vocabulary, which does not help to understand the processes taking place in the lexicon. All these processes violate the original linguistic systematicity in order to become the source of new systematic relations in the further stages of linguistic development.

In this paper, we attempt to test the non-arbitrary form-meaning hypothesis in the Old English lexicon. Statistical methods and semantically (conceptually) organized dataset of A Thesaurus of Old English allowed us to test the form-meaning association at the lexical level and conclude with certain assumptions a statistically significant form-meaning correlation.

The structure of the article is the following. In Introduction, we propose a hypothesis and set the objectives of the study; in Sections 2-4 we present the outline of the previous work and describe basic terminology; in Section 5 we outline the data and methods of the research and summarize the obtained results in Section 6. In Section 7 we outline the possible prospects of the current research and reveal its limitations.

2. Old English Lexicon

Old English is "the language spoken by the Germanic inhabitants of Britain" (5th–11th c.) in which prosaic and poetic texts were written (Fulk, 2014) and one of the earliest periods of language development (7th–11th c.). Old English prosaic texts include the translations of the Bible, Gospels, Psalter, Wulfstan's Homilies, Ælfric's works (religious texts); law codes, wills, and charters (legal texts); The Anglo-Saxon Chronicle (documentary prose); King Alfred's original compositions and translations from Latin (literary, philosophical, and didactic prose), historical works, and medical tracts. Poetry is represented by heroic and elegiac poems, religious and lyrical texts, magical and didactic poems, and riddles. Alliterative poetic texts, in particular, Beowulf, Genesis, Exodus, Cynewulf's poems Elene, Juliana, Andreas; Judith, The Dream of the Rood, The Wanderer, The Seafarer, Metrical Charms, and others have gained an important place in the history of Old English (Godden, 1992). There is a fairly large number of the surviving Old English texts. Nevertheless, religious texts are thematically dominant, which may impose constraints on the further analysis of the lexis.

The Old English lexicon, presented in A Thesaurus of Old English, was collected by lexicographers from the English texts of 7th–11th c. and

contains lexical layers of different chronological and etymological depth (Pollington, 1993). The lexicon includes the vocabulary of Latin/Greek origin, neutral groups of words of different genres, and a small amount of colloquial vocabulary (this layer of vocabulary cannot be fully selected due to the lack of speech fixation). Along with the vocabulary of different registers, the Old English lexicon includes purely poetic vocabulary found mainly in poetry and nowhere else (Barney, 1985).

The Old English lexicon and its various linguistic and stylistic aspects have been studied by many scholars (Kastovsky, 1992; J. J. Smith, 2009). Nevertheless, we are not aware of the works that study non-arbitrary form-meaning relationships in the Old English lexicon as a whole.

3.   Form-Meaning Relationships: Basic Terminology

Within the framework of non-arbitrary relationships in language, there are a number of terms that constitute the conceptual basis for this research. They include the terms linguistic sign, signifier, signified, lexeme, initial sound, grapheme, referential meaning, arbitrariness, iconicity, similarity, isomorphism, analogy, systematicity.

Linguistic signs represented by words or lexemes—the major units of vocabulary, are identified by the two components—the form (sounds/graphemes) and the content (meanings). In Saussure's theory these components are signifier and signified. Signifier is represented as a sound or graphical form of a word. The most relevant component of the form is an initial sound (phoneme). It is the smallest structural unit of language that carries important information since it is cognitively and positionally marked. An initial phoneme is connected with an initial grapheme—an orthographic representation of a sound. Signified is a word meaning—a quantum of sense revealing information about entities, processes, ideas, events in the world. Since there are a number of word meanings (lexical, grammatical, social, connotative, pragmatic, etc.) we discuss only conceptual meanings. They differ from other forms of meanings in that they refer to a cognitive content of a word. We can consider a conceptual meaning as the referential meaning of a word—an entry that is usually given in a dictionary.

'Signifier/signified' relations can be arbitrary or non-arbitrary (iconic). Arbitrariness does not hold any direct natural connection between the sign and its meaning and can be applied to the majority of linguistic signs. The principle of "whatever name you give to a thing is its right name" (Plato) became the fundamental principle of language whereby linguistic signs are considered arbitrary or conventional (F. de Saussure). Non-arbitrariness or iconicity involves the relations of similarity, isomorphism or analogy between some aspects of form and meaning.

The three terms are nearly identical in meaning, with similarity being a "one-to-one mapping of a Euclidean space onto itself"[2], isomorphism is "a correspondence (relation) between objects or systems of objects expressing the equality of their structures in some sense"[3] (cf. Givón (1985)), and "analogy between S and T is a one-to-one mapping between objects, properties, relations and functions in S and those in T".[4] Non-arbitrariness involving these relations enables us to predict word meanings. E.g., a word form (sounds/graphemes in *moo*) is associated with the referent (a mooing cow) and bears some resemblance to its semantics ('the sound produced by a cow').

In recent years, the term systematicity has been increasingly used. It is understood as the statistical form-meaning regularities found in "localized form-meaning patterns" or "across the lexicon as a whole" (Gutiérrez, Levy, and Bergen, 2016, p. 2379). The content and status of the term has not yet been clearly defined since it is applied both to arbitrary and non-arbitrary form-meaning relationships. It is explained as arbitrary statistically regular patterning of sounds (Dingemanse et al., 2015) or as "strong, non-negligible lexicon-wide non-arbitrariness" (Gutiérrez, Levy, and Bergen, 2016, p. 2380). The 'information' nature of the term is identified in Pimentel et al. (2019, p. 1752) that estimates the mutual information between the form and the meaning of a linguistic sign, i.e., the word-form/semantic distance/similarity. The authors point out that systematicity can be understood more broadly—as an umbrella term for all cases of regular patterning in language. In that case, systematic relations are manifested in obligatory grammatical/semantic oppositions, e.g., in grammatical categories of case/number or semantic oppositions of hypernymy/hyponymy.

4.   Non-Arbitrary Form-Meaning Relationships

Signifier/signified non-arbitrary relationships are studied in the theory of non-arbitrariness (iconicity). The research in the field of iconicity is extensive. Iconic and motivated signs are explored in language acquisition (Winter, Perlman, Perry, and Lupyan, 2017), cognitive (Wilcox, 2004) and neurolinguistic studies (Aryani, Jacobs, and Conrad, 2013); (Perniss and Vigliocco, 2014); (Monaghan, Shillcock, Christiansen, and Kirby, 2014), poetics (Tsur, 2002), etc. More and more studies focus on the evolutionary aspects of iconicity (Zlatev, Żywiczyński, and Wacewicz, 2020). In this and similar research, the authors define non-arbitrary form-meaning relationships, develop their classifications, and

---

2. http://encyclopediaofmath.org/index.php?title=Similarity&;oldid=31636

3. http://encyclopediaofmath.org/index.php?title=Isomorphism&oldid=21572

4. https://plato.stanford.edu/archives/spr2019/entries/reasoning-analogy/

disclose their nature. The terms (correspondence, equality, resemblance, congruence, equivalence, identity, analogy) that are used to interpret them are numerous, each reflecting subtle aspects of form-meaning association (mapping, correlation).

The definitions/mechanisms of non-arbitrary relationships vary depending on different types of linguistic signs. In onomatopoeia, non-arbitrary relationships are defined as the relations of low/high-order similarity or resemblance and structural similarity between form and meaning (Winter, Perlman, Perry, and Lupyan, 2017); (Dingemanse et al., 2015). In the former case, physical sensoriperceptual properties of the referent (usually emitted sounds) are imitated in the perceptual/graphical properties of the word form (sounds/graphemes) and are associated with some features of the lexical meaning of the word. The imitation of sensory properties refers to the low-order similarity where the attributes of the objects are compared. In the latter case, there is a relational similarity in the elements of structure (high-order similarity), where the relations between objects are compared, e.g., the sequential order of the events imitates the word order in a sentence (Haiman, 1985).

In sound symbolism, we also deal with the high-order similarity where the properties of abstract 'referents' are associated with the elements of a word form, and where the human—cognitive, motor, spatial, emotional, etc. experiences are symbolized by sounds. E.g., "high tones, vowels with high second formants (notably /i/), and high-frequency consonants are associated with high-frequency sounds, small size, sharpness, and rapid movement; low tones, vowels with low second formants (notably /u/), and low-frequency consonants are associated with low-frequency sounds, large size, and heavy, slow movements" (Hinton, Nichols, and Ohala, 1994, p. 10). According to Winter, Perlman, Perry, and Lupyan (2017), words referring to sensory domains (sound, sight, touch, taste, and smell) are more iconic than the words with abstract meanings. The differences between onomatopoeia/sound symbolism are reflected in the two types of iconicity—absolute or primary/relational or secondary correspondingly.

In phonaesthemes, the relations of systematicity are brought to the fore: they possess an initial cluster of phonemes which occurs regularly within a set of words. E.g., the words starting with *bl-, sn-, gl-, pr-,* etc. have some similarity of meaning, referring to such semantic classes as 'audible', 'perceptible', 'moving', etc. (Lvova, 2005). In this case, we are not concerned with a direct relationship between form and meaning but with a systematic correspondence between them. Such "regular mapping between aspects of form and function" (Dingemanse et al., 2015) features distributional regularities in different languages. Various scholars measuring sound/meaning similarity distance (Shillcock, Kirby, McDonald, and Brew, 2001); (Abramova and Fernández, 2016)

have arrived at the conclusion about the ubiquitous character of this phenomenon.

Most linguists study iconic signs within the framework of phonological iconicity. Our aim was to show the role and place of referential meanings in non-arbitrary form-meaning relationships in the history of the Old English language. Indeed, iconicity plays an important role in language development: according to Perniss and Vigliocco (2014), it "bridges between" language and human experiences and "support *referentiality*" (the ability of speakers to label objects and events in the processes of nomination) and "displacement" (the ability of linguistic signs to stand for the referents).

Within the course of language development iconicity erodes: a good example of erosion is the process of grammaticalization where the forms with more concrete meanings are superseded with the forms with more abstract meanings. E.g., the Old English verbs of existence *beon/wesan* used to nominate more concrete meanings of 'growth', 'biding', and 'dwelling' clearly capturing some iconic properties relevant for the speakers. In the course of time these meanings semantically eroded and later were replaced by the more abstract grammatical meaning of 'existence'. One more example of the erosion of iconicity is the language vocabulary: "with vocabulary growth, representational spaces comprising forms and meanings become more densely populated, thereby increasing the possibilities of confusion and ambiguity in the spoken forms of words, providing a selective pressure towards more arbitrary, more discriminable forms" (Dingemanse et al., 2015).

5.   Data and Methodology

Our research is performed on the basis of the Old English lexicon. Computational analysis in the field of diachronic linguistics is based on such data preprocessing as lemmatization, stemming, POS and semantic tagging, morphological and syntactic markup. Old English is a low-resource language. The limitation of Old English textual data and the absence of finished implementations for older languages preprocessing constitute considerable challenge. Due to the broad dialectal variation in Old English, there is a large number of orthographic variants, which makes it difficult to perform lemmatization, stemming and POS tagging (the existing Python implementations provide inaccurate results). Morphological and syntactic markup is available only for a small number of texts. Thus, the scope of our study was narrowed, and it was decided to focus on the analysis of A Thesaurus of Old English,[5] with the appli-

---

5. http://oldenglishthesaurus.arts.gla.ac.uk/

cation of statistical methods of the research. A Thesaurus is provided under the license.

The data comprise the lexemes from A Thesaurus sharing the initial consonantal graphemes (phonemes) <w>, <s>, <h>, and <p>. The lexemes with the initial graphemes <w>, <s>, <h> are the most frequent ones while the lexemes with the initial <p>, on the contrary, are the least frequent. After processing, the dataset for the analysis comprised the lexemes with four onsets: <w> (3359 words), <s> (4586 words), <h> (5210 words), and <p> (542 words) excluding entries with the compound lexemes written separately or with a hyphen (13,697 words in total). In A Thesaurus lexical meanings of Old English words are arranged into 18 conceptually organized semantic categories. The categories are: 1. The Physical World, 2. Life and Death, 3. Matter and Measurement, 4. Material Needs, 5. Existence, 6. Mental Faculties, 7. Opinion, 8. Emotion, 9. Language and Communication, 10. Possession, 11. Action and Utility, 12. Social interaction, 13. Peace and War, 14. Law and Order, 15. Property, 16. Religion, 17. Work, 18. Leisure. A certain semantic category is ascribed to each word, e.g., *wæþ* 'a ford' is given under the category 5. 'Existence'. We hypothesize that there might be some non-random distribution of semantic categories over the lexemes sharing initial graphemes.

Statistical methods are widely used in linguistic research. For the analysis of the distribution the most commonly used one is Pearson's Chi-square, in particular the Chi-square test and the coefficient of contingency. Despite frequent critical remarks on the application of Pearson's Chi-square in linguistics we consider it to be appropriate and reasonable for our dataset though with certain assumptions. Pearson's Chi-square attempts at making a conclusion whether a distribution observed is purely accidental, or whether it reflects a certain regularity. This statistical test is applied to a contingency table made up of the element frequency in the sampling unit to be compared with the total number of elements in this unit. The null hypothesis tested is that the difference between the element frequency is the result of random variations. The implementation of Pearson's Chi-square statistics is available in a number of software frequently used in corpus linguistics such as WordSmith Tools and AntConc but in our case the data preprocessing and Pearson's Chi-square statistics were performed within Python realization.

6.   Form-Meaning Hypothesis in a Thesaurus of Old English

For the analysis we calculated the frequency of lexemes starting with <w>, <s>, <h>, and <p> graphemes in every semantic category of A Thesaurus of Old English. The results are presented in Table 1.

TABLE 1. The distribution of the initial graphemes <w>, <s>, <p>, <h> in semantic categories

	<w>	<s>	<h>	<p>
The Physical World	193	335	247	25
Life and Death	553	809	932	109
Matter and Measurement	164	327	261	23
Material Needs	283	478	514	76
Existence	347	707	658	53
Mental Faculties	216	264	222	15
Opinion	146	64	167	26
Emotion	236	202	307	9
Language and Communication	118	157	69	14
Possession	17	34	67	0
Action and Utility	135	150	214	16
Social interaction	304	246	318	14
Peace and War	126	185	234	5
Law and Order	114	106	105	10
Property	75	64	66	11
Religion	228	275	698	75
Work	66	113	88	27
Leisure	38	70	43	34

Table 1 presents a contingency table of two categorical variables—the semantic category and the initial grapheme of the word. To check if the semantic categories are distributed among the lexemes with identical initial graphemes non-randomly, we applied the Chi-square test. The Chi-square statistics is commonly used for testing relationships between categorical variables. The Chi-square test measures the ratio of difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table, which enables to estimate the relationship between the variables. The null hypothesis of the Chi-square test is that the categorical variables are statistically independent. The null hypothesis will be recognized when the observed frequencies are less than the expected counts. In cases where the observed frequencies are greater than theoretically expected ones, the relationships between variables are statistically significant and present evidence for correspondence between them. The formula for the Chi-square test is:

$$\chi_c^2 = \sum_i \frac{(O_i - E_i)^2}{E_i},$$

where c = degrees of freedom, O = observed value(s), E = expected value(s).

The application of the Chi-square test to the resulting table shows the following statistics: χ^2 = 765.67, p = 3.637 × 10^{-128}, df = 51. The critical value of the Chi-square statistics with 50 < df < 55 varies from 67.51 up to 86.66 and more. Thus, the resulting statistics can be considered an example of non-random variation with the significant p-value. Thereby, we can conclude that the null hypothesis of the independence of variables can be rejected.

Statistics also offers to analyze the cell-wise contributions to the Chi-square to see where the evidence for the dependence is coming from. The contribution to the Chi-square quantifies the individual category contributions, i.e., how much of the total the Chi-square statistic is attributable to each category difference between observed and expected values. The contribution to the Chi-square is found by taking the squared difference between the observed count and the expected count then dividing by the expected count. The results of the contributions to the Chi-square are presented in Table 2. Larger values indicate a more substantial contribution to the overall Chi-square statistics.

TABLE 2. Contributions to the Chi Square for Old English initial <w>, <s>, <p>, <h> in 18 semantic categories

	<w>	<s>	<h>	<p>
The Physical World	0.05	16.84	10.79	1.41
Life and Death	2.23	0.02	0.35	2.01
Matter and Measurement	3.57	17.57	3.87	1.93
Material Needs	7.04	1.46	0.00	9.46
Existence	17.02	22.80	0.26	4.09
Mental Faculties	9.18	2.39	9.43	6.32
Opinion	22.52	37.28	1.23	6.31
Emotion	14.12	10.08	1.42	14.58
Language and Communication	10.39	11.51	33.13	0.00
Possession	5.00	0.81	10.67	3.75
Action and Utility	0.60	2.92	1.68	0.95
Social interaction	35.57	8.23	0.91	12.54
Peace and War	0.58	0.00	2.94	12.93
Law and Order	12.35	0.34	3.95	0.80
Property	9.16	0.96	3.18	0.70
Religion	23.04	54.23	93.18	11.84
Work	0.52	2.16	5.08	20.25
Leisure	1.20	1.05	10.64	97.10

To estimate the relationship of every initial grapheme to certain semantic category we computed the Chi-square test for alternative dis-

tribution contingency 2 × 2 tables compiled on the basis of Table 1 as follows:

TABLE 3. The alternative distribution contingency table

	religion	other semantic categories	total
<p>	75	467	542
other initials	1,201	11,954	13,155
total	1,276	12,421	13,697

In cases where the association between two variables can be statistically significant (the Chi-square and p-value results), the strength of this association can be very small. To decide whether the relationship between two categorical variables is important, the coefficient of contingency (Phi (or φ)) is computed. The Phi coefficient measures how closely the observations in rows and columns are associated with each other. The formula for the Phi coefficient is $\varphi = \sqrt{\frac{\chi^2}{n}}$, where n is the total number of observations.

Table 4 provides the Chi-square, p-value and Phi coefficient results only for the cases where the observed data exceed the expected figures. The authors reason such data elimination by focusing on positive correlation between the variables. The critical value of the Chi-square statistics with $df = 1$ is 3,84. The range of Phi values from 0.05 to 0.1 are considered to estimate moderate positive relationships.

The results of the Chi-square together with the coefficient of contingency and the contributions to the Chi-square for the words with the initial graphemes <w>, <s>, <p>, <h> turn out to be different for certain semantic categories. This testifies to the idea of the correspondence between the initial grapheme of the word and its semantics. The resulting values allow the authors to reject the null hypothesis as they manifest non-arbitrary association between the words starting with different onsets and semantic categories:

<w> 'Social interaction' (*wealdan* 'to rule', *werod* 'assembly'), 'Opinion' (*wlanc* 'arrogant', *wlitig* 'fair, noble'), 'Emotion' (*weorc* 'suffering', *wynsum* 'joyful'), 'Law and Order', 'Language and Communication', 'Mental Faculties', 'Property'.

<s> 'Existence' (*sǣlnes* 'a time of silence', *samod* 'without a break'), 'Matter and Measurement' (*samen* 'together', *sid* 'large'), 'The Physical World' (*seolfor* 'silver', *sicel* 'small stream').

<h> 'Religion' (*hālig* 'holy', *hals* 'salvation'), 'Possession' (*habban* 'to possess'), 'Peace and War.

TABLE 4. Chi Square p-value and Phi coefficient data for Old English initials <w>, <s>, <p>, <h> in 18 semantic categories

	<w> chi^2, p-value, φ	<s> chi^2, p-value, φ	<h> chi^2, p-value, φ	<p> chi^2, p-value, φ
The Physical World		26.87 2.17E-7 0.04		
Life and Death		0.045 0.83 0.002	0.69 0.4 0.007	2.57 0.1 0.01
Matter and Measurement		27.99 1.22E-7 0.05		
Material Needs		2.43 0.12 0.01	4.45E-5 0.99 5.70E-5	10.98 0.0009 0.03
Existence		39.33 3.59E-10 0.05		
Mental Faculties	12.83 0.0003 0.03	3.79 0.052 0.02		
Opinion	30.73 2.96E-8 0.05 / 5%		2.04 0.15 0.01	6.8 0.009 0.02
Emotion	19.79 8.63E-6 0.04		2.43 0.12 0.01	
Language and Communication	14.14 0.0001 0.03	17.76 2.51E-5 0.04		
Possession			17.74 2.53E-5 0.04	
Action and Utility	0.83 0.36 0.008		2.8 0.09 0.01	
Social interaction	50.36 1.28E-12 0.06			
Peace and War		0.006 0.94 0.0007	4.94 0.03 0.02	
Law and Order	16.76 4.23E-5 0.03			
Property	12.33 0.0004 0.03			0.74 0.39 0.007
Religion			165.79 6.13E-38 0.11	13.66 0.0002 0.03
Work		3.31 0.069 0.02		21.59 3.37E-6 0.04
Leisure		1.6 0.2 0.01		102.62 4.05E-24 0.08

<p> 'Leisure'(*pipe* 'a flute', *plega* 'play', 'sport'), 'Work' (*pāl* 'a spade', *pinn* 'nail'), 'Religion' (*preost* 'a priest', *postol* 'an apostle'), 'Material Needs', 'Opinion'.

7.   Conclusion

In our study we have made an attempt to test the form-meaning hypothesis in earlier stages of language. The outline of our research sketches the entire lexicon, not the localized phonaesthemic patterns which are mostly examined in the related studies. To explore such patterns the authors apply the methods of historical semantics for thorough linguistic analysis without reference to computational tools.

We analyzed the methods of computational linguistics and realized that in relation to older languages not all of them are suitable. We have decided to employ statistical methods for our research. Although there is a lot of criticism about the use of statistical methods in linguistics, the results are worth considering as the application of these methods is promising for identifying the regularities in language development. We computed the relationship of Old English words sharing consonantal initial graphemes (phonemes) and their conceptual (referential) meanings in the lexicon. We found out that there is a certain association between an initial grapheme and semantic category to which the word sharing this grapheme belongs. We have revealed a regularity that the distribution of semantic categories among the words starting with one initial grapheme differs from the distribution of semantic categories among the words starting with the other onsets. We assume that this regularity may take place by chance, but it is highly likely to be based on non-arbitrary form-meaning relations. This patterning may be random, but it may as well be determined by the iconic coding in language.

For a while, the form-meaning hypothesis was tested only on four initial graphemes in the Old English lexicon. We expect that for a larger number of graphemes the results would be different. In the future, we plan to expand the dataset with more initial graphemes and to undertake a full-scale research. Further explorations into the topic can be continued in the sphere of semiographemics since form-meaning relationships can be also traced in other semiotic modelling systems (symbolic writings and art) with specific implications for form-meaning hypothesis (Lotman, 2011).

References

Abramova, Ekaterina and Raquel Fernández (2016). "Questioning Arbitrariness in Language: a Data-Driven Study of Conventional Iconicity." In: *Proceedings of the NAACL-HLT, San Diego, California*, pp. 343–352.

Altmann, Gabriel (2004). "Script Complexity." In: *Glottometrics* 8, pp. 68–74.

Aryani, Arash, Arthur M. Jacobs, and Markus Conrad (2013). "Extracting Salient Sublexical Units from Written Texts: "Emophon," a Corpus-Based Approach to Phonological Iconicity." In: *Frontiers in Psychology* 4.654.

Atkinson, Mark, Gregory J. Mills, and Kenny Smith (2019). "Social Group Effects on the Emergence of Communicative Conventions and Language Complexity." In: *Journal of Language Evolution* 4.1, pp. 1–18.

Barney, Stephen A. (1985). *Word-Hoard: An Introduction to Old English Vocabulary (Yale Language Series)*. 2nd ed. Yale: Yale University Press.

Blasi, Damián E. et al. (2016). "Sound-Meaning Association Biases Evidenced across Thousands of Languages." In: *Proceedings of the National Academy of Sciences*. Vol. 113, pp. 10818–10823.

Cornell, Muriel (1981). "Varieties of Repetition in Old English Poetry. Especially in The Wanderer and The Seafarer." In: *Neophilologus* 65.2, pp. 292–307.

Dingemanse, Mark et al. (2015). "Arbitrariness, Iconicity, and Systematicity in Language." In: *Trends in Cognitive Sciences* 19.10, pp. 603–615.

Fulk, Robert D. (2014). *An Introductory Grammar of Old English with an Anthology of Readings*. 1st ed. Vol. 463. Medieval and Renaissance Texts and Studies. Tempe, Arizona: ACMRS Press.

Givón, Talmy (1985). "Iconicity, Isomorphism, and Non-Arbitrary Coding in Syntax." In: *Iconicity in Syntax. Proceedings of a Symposium on Iconicity in Syntax, Stanford, June 24–26, 1983*. Ed. by John Haiman. Amsterdam: Benjamins, pp. 187–219.

Godden, Malcolm R. (1992). "Literary Language." In: *The Cambridge History of the English Language. Vol. I: The Beginnings to 1066*. Ed. by Richard M. Hogg. Cambridge: Cambridge University Press, pp. 490–535.

Gutiérrez, E.Darío, Roger Levy, and Benjamin K. Bergen (2016). "Finding Non-Arbitrary Form-Meaning Systematicity Using String-Metric Learning for Kernel Regression." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Vol. 1. URL: https://www.aclweb.org/anthology/P16--1225.pdf.

Haiman, John (1985). *Natural Syntax: Iconicity and Erosion*. Cambridge: Cambridge University Press.

Hinton, Leanne, Johanna Nichols, and John J. Ohala, eds. (1994). *Sound Symbolism*. Cambridge University Press.

Jee, Hana, Monica Tamariz, and Richard Shillock (2018). "The Substructure of Phonics: The Visual Form of Letters and their Paradig-

matic English Pronunciation are Systematically Related." PsyArXiv https://psyarxiv.com/n85mb/.

Jespersen, Otto (1922). *Language: Its Nature, Development and Origin*. London: George Allen & Unwin.

Kastovsky, Dieter (1992). "Semantics and Vocabulary." In: *The Cambridge History of the English Language. Vol. I: The Beginnings to 1066*. Ed. by Richard M. Hogg. Cambridge: Cambridge University Press, pp. 290–408.

Lotman, Yuri (2011). "The Place of Art among Other Modelling Systems." In: *Sign Systems Studies* 39.2/4, pp. 249–270.

Lvova, Nadija L. (2005). "Semantic Functions of English Initial Consonant Clusters." In: *Glottometrics* 9, pp. 21–28.

Minkova, Donka (2003). *Alliteration and Sound Change in Early English*. Cambridge: Cambridge University Press.

Monaghan, Padraic et al. (2014). "How Arbitrary is Language?" In: *Philosophical Transactions of the Royal Society B—Biological Sciences* 369.

Perniss, Pamela and Gabriella Vigliocco (2014). "The Bridge of Iconicity: from a World of Experience to the Experience of Language." In: *Philosophical Transactions of the Royal Society B* 369.

Philps, Dennis (2008). "Submorphemic Iconicity in the Lexicon: a Diachronic Approach to English 'gn-words'." In: *Lexis* 2. URL: http://journals.openedition.org/lexis/728.

Pimentel, Tiago et al. (2019). "Meaning to Form: Measuring Systematicity as Information." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1751–1764.

Pollington, Stephen (1993). *Wordcraft: New English to Old English Dictionary and Thesaurus*. 2nd ed. Hereward, UK: Anglo-Saxon Books.

Shillcock, Richard et al. (2001). "Filled Pauses and their Status in the Mental Lexicon." In: *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech*, pp. 53–56.

Slaměníková, Tereza (2019). "On the Nature of Unmotivated Components in Modern Chinese Characters." In: *Proceedings of Graphemics in the 21st Century, Brest 2018. Ed. by Yannis Haralambous*. Ed. by Yannis Haralambous. Brest: Fluxus Editions, pp. 209–226.

Smith, Chris A. (2016). "Tracking Semantic Change in fl-monomorphemes in the Oxford English Dictionary." In: *Journal of Historical Linguistics* 6.

Smith, Jeremy J. (2009). *Old English: A Linguistic Introduction*. 1st ed. Cambridge: Cambridge University Press.

Tsur, Reuven (2002). "Aspects of Cognitive Poetics." In: *Cognitive Stylistics—Language and Cognition in Text Analysis*. Ed. by Elena Semino and Jonathan Culpeper. Amsterdam, Philadelphia: John Benjamins, pp. 279–318.

Wilcox, Sherman (2004). "Cognitive Iconicity: Conceptual Spaces, Meaning, and Gesture in Signed Language." In: *Cognitive Linguistics* 15.2, pp. 119–147.

Winter, Bodo et al. (2017). "Which Words are Most Iconic? Iconicity in English Sensory Words." In: *Interaction Studies* 18.3, pp. 430–451.

Youn, Hyejin et al. (2016). "On the Universal Structure of Human Lexical Semantics." In: *Proceedings of the National Academy of Sciences of the United States of America* 113.7, pp. 1766–1771.

Zlatev, Jordan, Przemysław Żywiczyński, and Sławomir Wacewicz (2020). "Pantomime as the Original Human-Specific Communicative System." In: *Journal of Language Evolution* 5.2, pp. 156–174.

Graphemic Complexity for the New Romance Phonemes in Italian

Some Reflections

Stefano Presutti

Abstract. The grapheme-phoneme correspondence (GPC) is considered the essential factor to classify the spelling consistency of world languages that also use a writing system to communicate. This paper focuses on cases of graphophonological inconsistency in a shallow writing system such as Italian. Particularly, this is an in-depth study of new Romance grapheme-phoneme correspondence complexity. This study attempts to explain why this inconsistency in Italian specifically involves these seven grapho-phonemes and does so by examining similar characteristics related to their historical development, which include some unsuccessful spelling reforms, phonological markedness, and language acquisition processes. In doing so, this paper examines the phenomenon of grapheme-phoneme correspondence consistency through original perspectives which therein provide a more complete picture of the possible motivations that led to these inconsistencies. The findings show that the surviving complexity related to these seven target consonants could indicate the effort that natives and non-natives should make to speak and write the standard language properly. Thus, at the grapho-phonological level, the etymological and national identity creation and preservation processes could be more important than the need to improve the language consistency.

1. Introduction

The correspondence between graphemes and phonemes of a language system is used as the main unit to classify the spelling consistency of all written languages. Particularly, by virtue of the Orthographic Depth Hypothesis (Frost, Katz, and Bentin, 1987; Katz and Frost, 1992), world languages have been classified in the last few decades according to their grapheme-phoneme correspondence (GPC) consistency degree along an axis with two extremes: orthographic depth and transparency.

Much of the research regarding GPC consistency has focused largely on English, a strong example of a language with high orthographic

Stefano Presutti   0000-0002-7700-5016
University of California Rome Center, Italy
E-mail: s.presutti@eapitaly.it

depth, to the detriment of other languages (Besse, 2007; Share, 2008; Ziegler, 2018). One of the main innovative elements of this study concerns the target language. Instead of analysing the grapho-phonological consistency of a deep writing system, and focusing instead on that of Italian, this paper has studied an inconsistent grapheme-phoneme correspondence present in a highly transparent writing system.

I sought to understand why Italian has a GPC inconsistency with specific new Romance phonemes, which were the hindmost institutionalized elements within the Italian phonological system, and why this has persisted as one of the primary unsolved problems of correspondence between sound and spelling.

Research on graphemic complexity has consistently preferred a synchronous study of the target language. In contrast, this paper has highlighted the grapheme-phoneme correspondence consistency through different perspectives, with the main goal of understanding what possible linguistic factors may affect the correspondence between graphemes and phonemes. Firstly, this study is one of the first attempts to clarify the history of target language grapho-phonemes in order to understand why there is GPC inconsistency in Italian despite its shallow orthography and the several attempted spelling reforms that have been made over time. Secondly, I have identified some compelling similarities between the seven inconsistent grapho-phonemes while considering their markedness degree, their diachronic period of sociopolitical acceptance into the grapho-phonological system, and their acquisition times in mother tongue (L1) and non-native (L2) contexts.

The remainder of this paper is organized as follows: in order to lay a foundation for the discussion of grapheme-phoneme correspondence consistency, especially in Italian, I discuss some grapho-phonological preliminaries. Following that, I describe the main characteristics of Italian graphematics, and then discuss how the grapho-phonological system has developed over time, particularly with regard to the seven new Romance target grapho-phonemes. I also show how several attempted spelling reforms failed. Moreover, I highlight similarities concerning the phonological markedness, the diachronic institutionalization, and the L1 and L2 learning processes. I distinctly describe the main feature of one of the grapho-phoneme targets: the palatal lateral approximant in Italian. The paper closes with a brief conclusion in which I summarise the findings and offer some suggestions for possible future proposals.

2.   The GPC Consistency

In order to delve deep into one of the main Italian GPC inconsistencies, it is essential to clarify the main theoretical terms used in this paper— namely, the *spelling-sound consistency* and *orthographic depth*.

The orthographic consistency of a language is characterized by a more or less exact correspondence of the sublexical units between the phonological and the graphematic systems. Orthographic consistency does not exclusively concern the GPC (grapheme-phoneme correspondence) or the PGC (phoneme-grapheme correspondence), but additionally the larger parts of a word, such as syllable, coda, and rhyme. Consistency—or transparency—can be measured in both directions: from sound to spelling or from spelling to sound. In addition, the concept of consistency can be partially complemented by that of orthographic depth. As Richlan recently identified, the orthographic depth is "the complexity, consistency, or transparency of grapheme-phoneme correspondences in written alphabetic language" (Richlan, 2014, p. 1). Therefore, this term only refers to the minimal units of language.

This concept has been mentioned in an array of research throughout the 1980s and early 1990s (Lukatela, Popadić, Ognjenović, and Turvey, 1980; Liberman, Liberman, Mattingly, and Shankweiler, 1980; Katz and Feldman, 1981; I. Y. Liberman, 1989; A. M. Liberman, 1992; Seidenberg, 1992), but it acquires a more definitive value when referencing the reliability of spoken and written language correspondences, apropos of the aforementioned Orthographic Depth Hypothesis. Notably, a shallow or transparent orthography will have a more direct spelling-sound correspondence, as is the case with Serbo-Croatian, while a deep orthography will exhibit a less direct one-to-one single grapheme-phoneme correspondence, as is the case with English (Katz and Frost, 1992).[1] However, the degree of transparency is often variable, even within the same language. Because of this, some languages, such as Spanish and Italian, have a high consistency from spelling to phonology, but also a lower consistency in the opposite direction (cf. for Spanish Landerl, 2006; for Italian Neef and Balestra, 2011).[2]

3.   Characteristics of Italian Writing System

I describe herein the main features of the target language of this study and, more specifically, the new grapho-phonemes introduced later in Italian.

---

1. It should be noted, however, that the measure of the complexity of the relationship between the orthographic and phonological systems of a language remains particularly complex and does not constitute a universal value. Indeed, each modality for classifying orthographic depth is based on a partial choice of rules to be considered (cf. Schmalz, Marinus, Coltheart, and Castles, 2015; Ziegler, 2018).

2. As a general rule, we can say that when there is a difference in consistency between spelling and sound, phoneme-to-grapheme correspondence tends to be less transparent than grapheme-to-phoneme correspondence (Cook and Bassetti, 2005, pp. 9–10).

When considering the primary rules of Italian GPC and comparing them to those of other European languages (cf. Table 1), Italian emerges as being more transparent than others, such as English or French, because it has fewer rules for phoneme-grapheme correspondence. Thus, Italian has a shallow writing system to the extent that it is mostly written as it is pronounced (Maraschio, 1993).

TABLE 1.   Measures of complexity and unpredictability for Dutch, English, French, German and Italian (Schmalz, Marinus, Coltheart, and Castles, 2015)[3]

	Dutch	English	French	German	Italian
Total number of rules	104	226	340	130	59
Single-letter rules	51	38	46	44	19
	(49.0%)	(16.9%)	(13.5%)	(33.8%)	(32.2%)
Multi-letter rules	42	161	218	55	8
	(40.4%)	(71.2%)	(64.1%)	(42.3%)	(13.6%)
Context-sensitive rules	11	27	76	31	32
	(10.6%)	(11.9%)	(22.4%)	(23.8%)	(54.2%)

As illustrated in Table 2, there are some cases in contemporary Italian in which a phoneme is represented by a complex grapheme, wherein a grapheme changes depending on the context. Additionally, a few complex correspondence rules with some graphemes also exist. In each of these circumstances, with the exception of double consonants (in the fifth line) and vowel ortho-epic characteristics (in the third line), the same graphic signs are always used: <g> (also used for <gl> and <gn>), <z>, <s>, <c>, and the diacritical letter <i>. The following paragraph shows how all of them have been adopted to represent the new Romance phonemes introduced in the Italian phonological system.

3.1.   Complex Graphemes of New Romance Phonemes

The elements institutionalized last in the Italian grapho-phonological system consist of seven consonants, as reported in Table 3: four dental affricates (alveolars /ts - dz/ and prepalatals /tʃ - dʒ/), the prepalatal fricative /ʃ/, and the palatals (the nasal /ɲ/ and lateral /ʎ/).

The alveolar affricates /ts/ and /dz/ are represented by the same single-letter <z>, which creates a homographic situation, while the

---

3. Measures of complexity and unpredictability are based on the Dual-Route Cascaded Model (or DRC, see Ziegler, Perry, and Coltheart, 2000; Rastle and Coltheart, 1998; Paap and Noel, 1991). For the DRC model, the numbers represent the number of rules of each type, and the percentage out of the total number of rules in brackets.

TABLE 2. Graphemic Complexity for the New Romance Phonemes in Italian (Neef and Balestra, 2011)

Number of letters	21	
Fixed letter combinations	3	(<gl>,<gn>, <sc>)
Undetermined	5	(<e>, <i>, <o>, <s>, <z>)
Context-dependent	6	(<c>, <g>, <gl>, <i>, <s>, <sc>)
Inherently ordered	13	(<b>, <c>, <d>,<f>, <g>, <l>, <m>, <n>, <p>, <r>, <s>, <t>, <v>)
Complex correspondence rules	4	(<c>, <g>, <i>, <s>)

TABLE 3. Graphemes representing the seven new Romance consonants in Italian

	Followed by /a, o, u/	Followed by /e/	Followed by /i/
/ts/	z	z	z
/dz/	z	z	z
/tʃ/	ci	c	c
/dʒ/	gi	g	g
/ʃ/	sci	sc	sc
/ɲ/	gn	gn	gn
/ʎ/	gli	gli	gl

palatal nasal is always represented by a digraph. All other cases present heterographic situations: the affricates' graphemes are both a single-letter grapheme and a digraph, while the prepalatal /ʃ/ and the palatals /ɲ - ʎ/ are always written with complex graphemes (digraphs or tri-graphs). Furthermore, the graphic representation of the affricates and palatals changes depending on the following vowel. Finally, the spelling of the three prepalatals /tʃ - dʒ - ʃ/ is somewhat ambiguous (see Table 4) because, in the Italian writing system, they can be easily confused with velar stops and the consonant sequence /sk/. In fact, these phonemes are graphically differentiated from each other simply by using the dia-critical letters <i> and <h>.

TABLE 4. Graphemes representing Italian prepalatals (2 affricates and 1 frica-tive), velar stops and a consonant cluster (fricative + stop)

/tʃ/	c - ci	/k/	c - ch
/dʒ/	g - gi	/g/	g - gh
/ʃ/	sc - sci	/sk/	sc - sch

4.   Diachronic Development

Italian is a Romance language spoken today primarily in Italy, and it is derived from the vernacular spoken in Florence in the fourteenth century. Like other contemporary Romance languages, such as French, Spanish, Portuguese, and Romanian, Italian is a linguistic continuation of Latin. At the beginning of its language development path, from the tenth to the fifteenth century, a linguistic pastiche indicated the lack of a real linguistic border between Latin and Romance Italian. In fact, if we compare the Romance languages according to their diachronic typology, the Florentine-based Italian represents one of the linguistic systems that are the least distant from the initial Latin matrix, and therefore more conservative compared to Romance languages such as French and Romanian, whose evolutions are the most marked (Banniard, 2008).

Until the sixteenth century, Latin remained as the main written language in the Italian states. For a long time, the Romance language struggled to have enough distance from its language of origin, an attribute vital to create a new sociolinguistic identity for the same community. Although a number of Florentine and non-Florentine intellectuals eventually succeeded in making Italian independent of Latin, their initial proximity is the main cause of the GPC inconsistencies still present in Italian today. Particularly in the first period of time, different structures of language were influenced by other Romance languages, especially French and Provençal, even at the grapho-phonological level. They were uniquely appreciated and used in the northern regions of Italy, close to the Alpine border, but their success and prestige also had considerable influence on Italo-Romance languages geographically more distant, such as the formal Sicilian used by several poets at the court of the Holy Roman Emperor Frederick II. In the first centuries of Italian diachronic development, there were numerous spelling variations at the individual, local, and regional levels (cf. Cornagliotti, 1988; Maraschio, 1993; Presutti, 2019). Within the same text, it was even possible to use multiple alphabets apart from Latin, such as Hebrew, Greek, and Arabic (cf. Coluccia, 2002).

In the sixteenth and seventeenth centuries, a substantial debate developed around the creation of a single standard language, both oral and written, for all Italo-Romance communities. This common objective of most Italian scholars was supported by the printing revolution as well. For that reason, a common writing system was institutionalized, fixed by a set of rules. The written language played a pivotal role in the standard oral language development. Writing was considered the basis for speech, serving as its model and its point of departure, rather than one of its subsequent steps (a trajectory unlike those of other European languages).

After that period of time, the standardized version of the Italian writing system did not change until recently, despite linguistic and political unification and the beginning of the mass literacy phase.

4.1. Attempted Spelling Reforms

To further explicate the previous section, I present an in-depth analysis of the language reforms which tried to improve the Italian GPC inconsistency throughout the centuries, from its origins to the present day.

Distinguished scholars have proposed many attempted spelling reforms over time; however, there were two particular types of spelling reformers: the etymologists, who wanted to reduce the distance of the Latin roots of words, and the phoneticians, who wanted to improve the Romance grapheme-phoneme correspondence and to accelerate the written comprehension and production. In these ways, they tried to solve the homographic and heterographic complexity derived from the use of the Latin alphabetical system, the language of origin with a different phonological system. Despite their attempts, none of the proposals to change spelling were accepted. It is for this reason that the Italian language still presents the same spelling-sound inconsistencies of the sixteenth century, the period in which its spelling was standardized. This was not so much due to a conservative tendency of the written language in relation to the oral one[4], but rather to a form of inertia in the Italian alphabetical system.[5]

Below, I give three examples of spelling reform proposals in Italian.[6] The first one was suggested in 1435 by the Florentine intellectual and architect Leon Battista Alberti who was the same author of the first Italian grammar. Alberti suggested listing the alphabetical letters (standard and new graphemes) in a different order based on the graphic complexity: from the easiest to write to the most difficult (cf. Fig. 1). However, his spelling reform proposal was incomplete because he did not consider all the phonemes present in Italian, and thus it did not solve the GPC inconsistency.

During the standardization period between the sixteenth and seventeenth centuries, many literates such as Trissino, Bartoli, Tolomei, and Fiorenzuola, attempted to improve the grapheme-phoneme correspondence consistency. The Italian spelling institutionalization period was

---

4. For centuries, the development of the Italian writing system followed a different path from that of pronunciation.

5. There have been several forms of resistance: socio-educational, economic and aesthetic. For further details, see Maraschio, 1993.

6. For a more detailed description, see Maraschio, 1992b; Presutti, 2019.
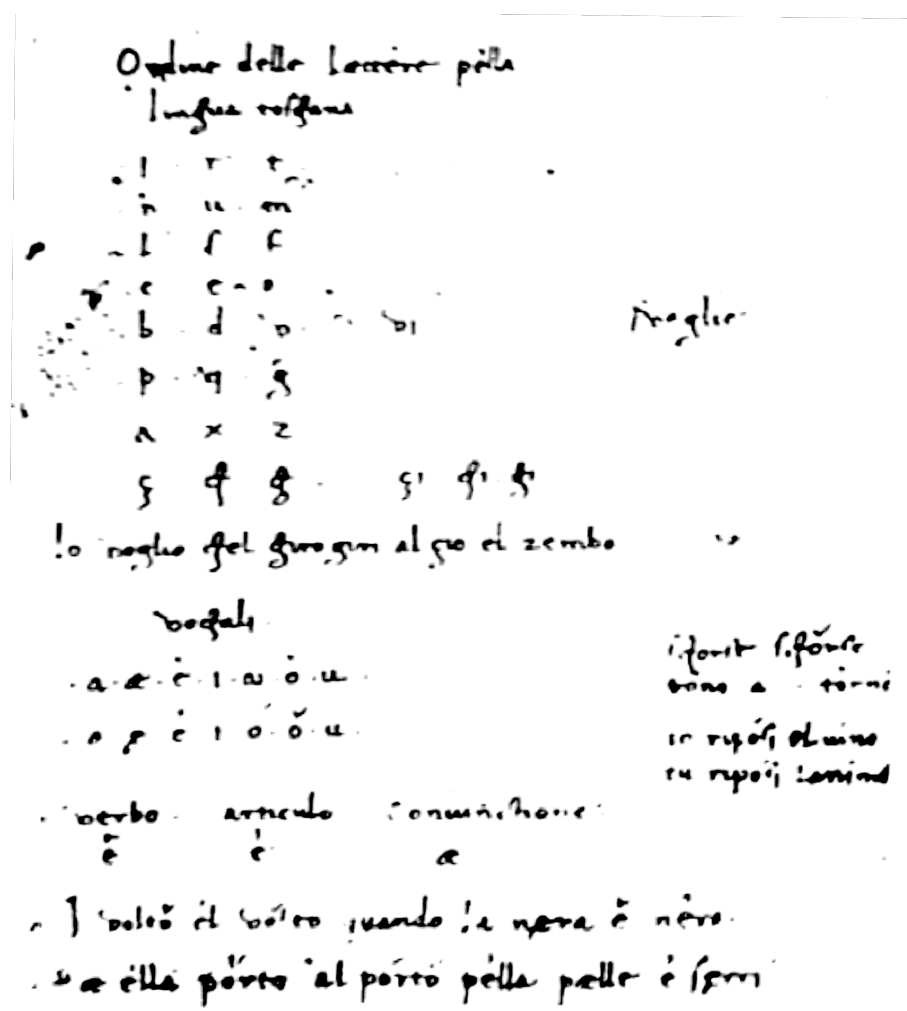
FIGURE 1. Alphabetical reform proposed by Leon Battista Alberti (Gorni, 2012)

accompanied by the same enthusiasm and several radical reform proposals. In fact, many grammarians criticized the use of the Latin alphabet as inadequate for Italian phonology, which led to debates about possible changes to the old writing system.

   Another example of spelling reform was proposed by Giorgio Bartoli in 1584 (cf. Fig. 2). According to him, the main aim of spelling was to maintain perfect clarity in the one-to-one relationship between phoneme and simple grapheme, thus avoiding homographic and heterographic solutions and the use of digraphs or trigraphs. He urged

for an alphabet comprising thirty-five phonemes and corresponding graphemes. In reality, his proposal did not present a perfect bilateral correspondence for the seven target consonants of this paper. Instead, the two prepalatal affricates /tʃ/ and /dʒ/ as well as the velar stops /k/ and /g/ were represented by two or three graphemes. Nonetheless, this was one of the best examples of attempted consistent bi-directionality between graphemes and phonemes in Italian. Yet because his proposal was not welcomed by the politicians, grammarians, or other intellectuals, the Italian writing system, again, did not change.

a	animo		1	i	io		19
b	bontà		2	j	jerico	ierico	20
c	cera		3	l	leone		21
ꝗ	ꝗane	cane	4	m	mare		22
q	diꝗo	dico	5	n	nero		23
h	pehe	pesce	6	ɲ	veɲto	vento	24
ɦ	peɦe	pece	7	o	moro		25
ɕ	ɕaue	chiaue	8	ɔ	ɔra	ora	26
d	dono		9	p	pane		27
e	il mele		10	r	riua		28
ɛ	mɛlo	melo	11	s	casa		29
f	fiore		12	ʃ	roʃa	rosa	30
ʒ	ʒente		13	t	terra		31
g	girlanda	ghirlanda	14	u	umile		32
ʎ	maʎo	maglio	15	v	via	uia	33
ɢ	maɢo	magno	16	ʒ	ʒelo		34
ɖ	ɖaccio	ghiaccio	17	z	zana	zana	35
ʝ	aʝo	agio	18				

FIGURE 2. Alphabetical reform proposed by Giorgio Bartoli (Maraschio, 1992b)

Following this period of time, the interest in innovating the writing system was considerably reduced. Only the Italian political unification in the nineteenth century, and its related socio-political changes, again fueled the importance of improving the Italian spelling learning process in school education programs.

A third and final example of attempted spelling reform was suggested by the politician and glottologist Goidanich in 1910. Fig. 3 exhibits

five out of seven reported target graphemes and excludes the alveolar affricates. He created one distinctive sign for each of the seven new Romance consonants, basically merging lines and curves of digraphs and trigraphs into just one single sign. Thus, he ultimately resolved the biunivocity of these seven new consonants. Yet again, however, Goidanich's and other similar orthographic reforms did not succeed in radically changing the writing system's inconsistencies.
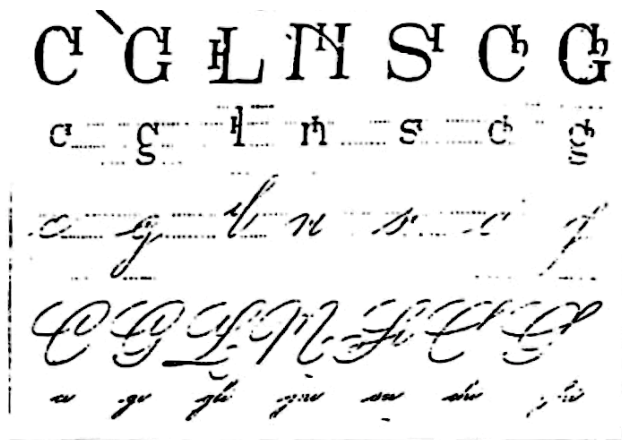
FIGURE 3. Alphabetical reform proposed by Pier Gabriele Goidanich (Goidanich, 1910)

5.   New Romance Grapho-Phoneme Similarities

After describing the main diachronic steps of the seven target consonants' inconsistency, I discuss the similarities between them in this chapter.

I noticed previously (cf. section 3.1) that from an orthographic point of view, they present a complex graphemic unit. From a phonetic point of view, they have a high markedness because they are less common than other sounds in world language phonetic classifications (Ladefoged and Maddieson, 1996; Maddieson, 1984).

When also considering the phonological hierarchy proposed by Jakobson (1968), the appearance of phonemes in language learning follows a precise universal hierarchy, dividing the vocal tract into smaller and smaller sections in order to create the phonemic identified by Trubetzkoy (1971). If compared with other consonants such as stops, the

seven new Romance target phonemes are more difficult to produce (and to hear, according to the Quantal theory by Stevens, 1989[7]). Thus, they appear later or do not appear at all in the distinctive phonetic process that is the basis of the phonological system of each language. Considering the Italian learning process specifically, these phonemes are produced last both in mother tongue and second language context.[8] In Table 5, I report the results of a recent experiment conducted by Italian speech therapists on the phonological development of Italian-speaking children (Tresoldi et al., 2018). They measured the average age of customary production, acquisition, and mastery of Italian consonants in an L1 context. The results, based on a large sample of participants aged 3 to 7 years old and representative of different geographical areas, showed that consonants such as plosives were mastered early by Italian children, while our seven target phonemes were the hindmost acquired segments.

To consider the writing as well, among the most common spelling mistakes made by native children and illiterates are again the complex graphemes representing these seven Romance phonemes. In Table 6, there are some examples of common spelling errors collected by Dardano (1993) and Tressoldi, Cornoldi, and Re (2017). The index of a recent Italian L2 textbook, reported in Fig. 4, show the same situation found in the mother tongue context. The textbook *Domani 1* (Guastalla and Naddeo, 2010) was targeted to beginners, particularly adult non-native speakers. First of all, the Italian second language pronunciation is taught toward the alphabet instead of the phonemes' acquisition; this spelling dominant approach is detrimental for learning all phonemes not represented by one single letter. Furthermore, complex graphemes are avoided, and they are only proposed later in various stages (particularly on chapters 3, 9 and 11; cf. Fig. 4). In the beginning, non-native students who study with this textbook can learn the grapheme-phoneme bilateral correspondences represented by the alphabet; however, they have to wait several units before learning the complex graphemes. Notably, in the third unit, they start using the contrast between affricates and velar stops, in the ninth unit the fricative, and finally, in the eleventh unit, they learn the nasal and lateral palatals.

In summary, the phonological learning order follows a precise hierarchy: in the beginning, it seeks the GPC consistency with the alphabet acquisition, and then it follows word frequency and phonological markedness parameters for the remaining phonemes represented by digraphs and trigraphs.

---

7. The notions of ease of articulation and auditory distinctiveness as influences on the phonetic structure of languages were suggested also by Martinet (1964), Lindblom (1990), Lindblom and Maddieson (1988).

8. In addition, in chapter 4 we noticed that they were institutionalized late in the standard language.

▸ **indice**

comunicazione	grammatica	lessico	testi scritti e *orali*	cultura

unità 0 | come ti chiami? pagina 11

comunicazione	grammatica	lessico	testi scritti e *orali*	cultura
• Chiedere e dire il nome • Le espressioni *Che significa?, Come si scrive?, Come scusa?* • Le operazioni aritmetiche • Salutare	• L'alfabeto • I numeri da 1 a 30 • Il verbo *chiamarsi* (io, tu, lui/lei)	• I nomi propri • I saluti	• *I saluti* • *L'alfabeto*	• Modi per salutarsi • Nomi propri più diffusi

unità 1 | di dove sei? pagina 18

comunicazione	grammatica	lessico	testi scritti e *orali*	cultura
• Chiedere e dire la provenienza e la destinazione	• I verbi *andare* e *essere* (io, tu, lui/lei)	• Le espressioni *grazie, prego, scusa* • Stazione e aeroporto	• Scritte in luoghi pubblici • *Annunci alla stazione* • *Dialogo in treno*	• Città italiane • Fare conoscenza

unità 2 | mi dai il tuo numero? pagina 23

comunicazione	grammatica	lessico	testi scritti e *orali*	cultura
• Chiedere e dare il numero di telefono • Chiedere l'età	• Il verbo *avere* (io, tu, lui/lei) • I numeri da 0 a 100	• Dati anagrafici	• *Dialogo in treno*	• Scambiare i dati anagrafici

unità 3 | tutti in piazza! pagina 26

comunicazione	grammatica	lessico	testi scritti e *orali*	cultura
• Aprire una telefonata • Concordare il luogo di un appuntamento • L'espressione *Come si dice in italiano...?*	• I nomi	• Luoghi della città	• Volantino • *Dialogo in treno*	• Ecologia • Città italiane

▸ Storia a fumetti Episodio 1   pagina 30
▸ Fonetica I suoni [k] e [tʃ] / I suoni [g] e [dʒ]   pagina 32

unità 9 | al bar pagina 69

comunicazione	grammatica	lessico	testi scritti e *orali*	cultura
• Salutare in modo informale e formale • Richiamare l'attenzione di qualcuno in modo informale e formale • Chiedere e dire il prezzo	• Gli articoli indeterminativi	• Cibi e bevande al bar • Tipi di acqua	• *Dialogo al bar* • Menù	• Andare al bar

▸ Storia a fumetti Episodio 3   pagina 74
▸ Fonetica I suoni [sk] e [ʃ] / Le doppie   pagina 76

unità 10 | la mia giornata pagina 78

comunicazione	grammatica	lessico	testi scritti e *orali*	cultura
• Dire a che ora si fa una cosa • Dire in che momento della giornata si fa una cosa	• I verbi riflessivi • *Anche / Neanche* • Gli articoli con i giorni della settimana • I possessivi	• Azioni quotidiane • Gli avverbi di frequenza • I giorni della settimana	• Fumetto umoristico • Forum	• Il bagno in Italia

unità 11 | in famiglia pagina 84

comunicazione	grammatica	lessico	testi scritti e *orali*	cultura
• Parlare della propria famiglia • Esprimere accordo o disaccordo • Fare una proposta e accettare • Incoraggiare • Introdurre un nuovo discorso	• Gli aggettivi possessivi e i nomi di parentela • *C'è / Ci sono* • I numeri dopo 1.000	• Nomi di parentela • Oggetti personali	• *Dialogo a casa* • Lettere ad un giornale	• La famiglia italiana

▸ Storia a fumetti Episodio 4   pagina 90
▸ Fonetica I suoni [ʎ] e [ɲ] / Le doppie   pagina 92

FIGURE 4. Index of the Italian L2 textbook Domani 1 (Guastalla and Naddeo, 2010)

TABLE 5. Age of acquisition of Italian phonemes (years; months) (Tresoldi et al., 2018)

	Age of customary production (≥ 50 %)	Acquisition age (≥ 75 %)	Mastery (≥ 90 %)
[p]			≤ 3; 0
[t]			≤ 3; 0
[m]			≤ 3; 0
[n]			≤ 3; 0
[b]		≤ 3; 0	3; 6
[l]		≤ 3; 0	3; 6
[k]		3; 6	4; 0
[d]		≤ 3; 0	4; 0
[f]		≤ 3; 0	4; 0
[v]	3; 6	4; 0	4; 6
[g]	≤ 3; 0	4; 0	4; 6
[ɲ]	3; 6	4; 0	5; 6
[dʒ]	≤ 3; 0	4; 0	5; 6
[ʃ]	≤ 3; 0	4; 6	5; 6
[tʃ]	≤ 3; 0	4; 0	6; 0
[r]	4; 0	4; 6	6; 0
[z]	≤ 3; 0	3; 6	6; 6
[ts]		6; 0	6; 6
[dz]	3; 6	5; 6	7; 0
[ʎ]	5; 0	6; 0	7; 0
[s]	≤ 3; 0	5; 6	7; 6

TABLE 6. Illiterates and children's wrong spelling examples

<g> - <gi>	litigare > lit**gi**are
<c> - <ci>	arance > aran**cie**
<gli> - <gl>	figlia > fi**gl**a
<gn> - <gni>	montagna > monta**gni**a

6.   An Example: the Palatal Lateral Approximant

Now I focus on one of these seven new Romance phonemes in Italian: the palatal lateral approximant /ʎ/. A more in-depth examination of one of the target elements can help to better understand the possible and diversified motivations that led to the contemporary grapheme-phoneme correspondence inconsistency. Moreover, the palatal lateral's institutionalization and survival in Italian could serve as an ideal example of how the processes of etymological conservation and national identity creation can be more important than GPC consistency improvement.

   This New Romance phoneme was mainly used in the Florentine dialect—which was the most influential Italo-Romance dialect and the

basis of standard Italian—and in formal contexts by Italian literates and nobles.[9] The palatal lateral presents an extremely high phonological markedness among world languages (Maddieson, 1984). When compared with the other more frequent Italian lateral, the alveolar /l/, the comprehension and production of the palatal are more complex (cf. Figs. 5 and 6). From both an acoustic and articulatory point of view, it can be easily misunderstood—by natives and others—with other sounds present in the Italian phonetic panorama such as the yod and the phonemic group /lj/ (Bladon and Carbonaro, 1978; Oliveira et al., 2016).
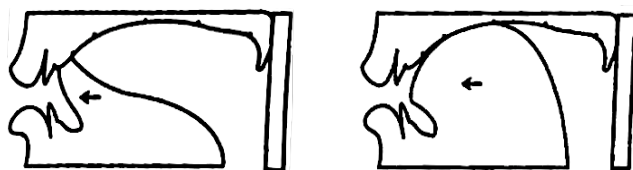
FIGURE 5. Sagittal sections of the lateral alveolar /l/ and palatal /ʎ/ (Canepari, 2004)
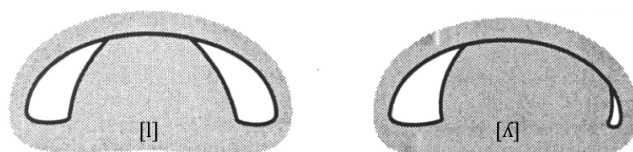
FIGURE 6. Transverse sections of the oral cavity: lateral articulation of /l/ and unilateral of /ʎ/ (Canepari, 2004)

To refer to the previous study reported in Table 5 (cf. Tresoldi et al., 2018), the palatal lateral is one of the last learnt phonemes by a native child, mastered at 6-7 years instead of 3-4 years like most of the other ones.

With regard to its spelling, the palatal lateral is still represented today by a digraph or a trigraph, depending on the following vowel (cf. Table 7). Before the standardized version of the sixteenth and seventeenth century, this phoneme was represented by a high number of graphemes (the most common ones are exhibited in Table 8). Additionally, there were several spelling alternatives proposed by scholar reformers over

---

9. For further details of the diachronic development of the palatal lateral in Italian, see Presutti (2019).

time (some of them are represented in Fig. 7). As previously mentioned, all of them were ignored by the political institutions.

TABLE 7. Current graphemes of

<gli>	<gl>
followed by /a, e, o/	followed by /i/
i.e., *foglia, foglie, foglio*	*fogli*

TABLE 8. Graphemes representing the over time (in the word *moglie*, wife)

	<lli>	<gl>	<lgl>	<lg>
i.e., *molie*	*mollie*	*mogle*	*molgle*	*molge*
<lgi>	<lgli>	<ll>	<lh>	<lhy>
i.e., *molgie*	*molglie*	*mullere*	*mulhere*	*mulhyere*

(a)   (b)   (c)
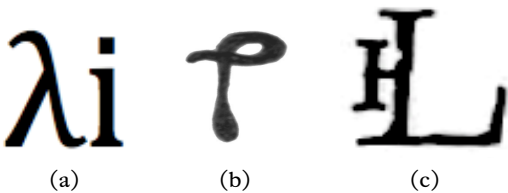
FIGURE 7. Some graphic alternatives proposed over time by spelling reformers such as (a) Tolomei in 1525, (b) Bartoli in 1584 and (c) Goidanich in 1910

7.   Conclusion

This paper has presented some insights into grapheme-phoneme correspondence inconsistencies in a highly transparent spelling system such as Italian, particularly apropos of the seven last institutionalized elements in the Italian phonological system. These target consonants are among the most difficult grapho-phonemes to be learnt and mastered. This paper has demonstrated that the reasons why they represent the main grapheme-phoneme correspondence inconsistency in Italian are strongly linked with the diachronic language development and with similar characteristics of phonological markedness and learning.

In general, the Italian writing system's strong resilience to change and improvement can be explained with the positive concept of "relative imperfection" which, in an original way, resolves the complex linguistic-identity stratification of Italian-speaking communities. On one hand, this national writing system was considered a stable and common communication tool for a highly heterogeneous population. On the other hand, it was considered flexible as it was able to accept all regional and local dialectal oscillations. Those two features allowed the Italian people to maintain their multiple linguistic identities. From a semantic point of view, the oral and written complexity of these seven New Romance grapho-phonemes could represent the effort that even native Italian speakers should make in order to speak and write the dominant language correctly. Thus, future studies of the GPC of other languages should consider the importance of the etymological and national identity creation and preservation processes. In truth, they are vital to understanding grapheme-phoneme correspondence inconsistency. In conclusion, this paper has offered alternative paths to explain GPC inconsistencies in a shallow language such as Italian. It has done so in the hope that the diachronic language development, phonological markedness, and grapho-phonological acquisition processes will be considered for further academic discussion concerning the GPC consistency of deep and shallow writing systems.

References

Banniard, Michel (2008). *Du latin aux langues romanes*. Paris: Armand Colin.

Besse, Anne-Sophie (2007). "Caractéristiques des langues et apprentissage de la lecture en langue première et en français langue seconde: perspective évolutive et comparative entre l'arabe et le portugais." PhD thesis. Université Rennes 2.

Bladon, R. A. W. and Emanuela Carbonaro (1978). "Lateral consonants in Italian." In: *Journal of Italian Linguistics* 3.1, pp. 43–54.

Canepari, Luciano (2004). *Il MaPI. Manuale di pronuncia italiana*. Bologna: Zanichelli.

Coluccia, Rosario (2002). *Scripta mane(n)t. Studi sulla grafia dell'italiano*. Galatina: Congedo.

Cook, Vivian J. and Benedetta Bassetti (2005). "An introduction to researching Second Language Writing Systems." In: *Second language writing systems*. Clevedon, UK: Multilingual Matters, pp. 1–67.

Cornagliotti, Anna (1988). "Geschichte der Verschriftung / Lingua e scrittura." In: *Lexicon der Romanistischen Linguistik Italienisch, Korisch, Sardisch (LRL)*. Vol. IV. Tübingen: Max Niemeyer Verlag, pp. 379–392.

Dardano, Maurizio (1993). "Profilo dell'italiano contemporaneo." In: *Storia della lingua italiana. Scritto e parlato*. Vol. II. Torino: Einaudi, pp. 405–430.

Frost, Ram, Leonard Katz, and Shlomo Bentin (1987). "Strategies for visual word recognition and orthographical depth: a multilingual comparison." In: *Journal of Experimental Psychology. Human Perception and Performance* 13.1, pp. 104–115.

Goidanich, Pier Gabriele (1910). *Sul perfezionamento del'ortografia nazionale e per la fondazione di una Società ortografica Italiana*. Modena: A. F. Formíggini.

Gorni, Guglielmo (2012). *Leon Battista Alberti. Poeta, artista, camaleonte*. Ed. by Allegretti Paola. Roma: Edizioni di storia e letteratura.

Guastalla, Carlo and Ciro M. Naddeo (2010). *Domani 1*. Firenze: ALMA Edizioni.

Jakobson, Roman (1968). *Child Language, Aphasia and Phonological Universals*. The Hague: Mouton.

Katz, Leonard and Laurie B. Feldman (1981). "Linguistic coding in word recognition." In: *Interactive processes in reading, Hillsdale*. Ed. by A.M. Lesgold and A. Perfetti. Hillsdale, NJ: Lawrence Erlbaum, pp. 85–105.

Katz, Leonard and Ram Frost (1992). "Reading in different orthographies: The orthographic depth hypothesis." In: *Orthography, phonology, morphology and meaning*. Amsterdam, NL: Elsevier, pp. 67–84.

Ladefoged, Peter and Ian Maddieson (1996). *The Sounds of the World's Languages*. Oxford-Cambridge: Blackwell.

Landerl, K. (2006). "Reading acquisition in different orthographies: Evidence from direct comparisons." In: *Handbook of orthography and literacy*. Mahwah, NJ: Lawrence Erlbaum, pp. 513–530.

Liberman, Alvin M. (1992). "The Relation of Speech to Reading and Writing." In: *Orthography, phonology, Morphology, and Meaning*. Vol. 94. Advances in Psychology. Amsterdam: North-Holland, pp. 167–178.

Liberman, Isabelle Y. (1989). "Phonology and Beginning Reading Revisited." In: *Haskins Laboratories Status Report on Speech Research* 105/106, pp. 1–8.

Liberman, Isabelle Y. et al. (1980). "Orthography and the Beginning Reader." In: *Orthography, Reading, and Dyslexia*. Ed. by J.F. Kavanagh and R.L. Venezky. Baltimore: University Park Press, pp. 137–153.

Lindblom, Björn (1990). "Explaining Phonetic Variation: A Sketch of the H&H Theory." In: *Speech Production and Speech Modelling*. Ed. by William J. Hardcastle and Alain Marchal. NATO ASI Series. Dordrecht: Springer Netherlands, pp. 403–439.

Lindblom, Björn and Ian Maddieson (1988). "Phonetic universals in consonant systems." In: *Language, Speech and Mind*. Ed. by C. Li and Hyman L. M. London: Routledge, pp. 62–78.

Lukatela, Georgiie et al. (1980). "Lexical decision in a phonologically shallow orthography." In: *Memory & Cognition* 8.2, pp. 124–132.

Maddieson, Ian (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.

Maraschio, Nicoletta (1992a). *Grafia e ortografia: formazione, codificazione, diffusione del sistema grafico italiano*. Firenze.

———— ed. (1992b). *Trattati di fonetica del Cinquecento*. Firenze.

———— (1993). "Grafia e ortografia: evoluzione e codificazione." In: *Storia della lingua italiana. I luoghi della codificazione*. Ed. by Luca Serianni and Pietro Trifone. Vol. I. Torino: Einaudi, pp. 139–227.

Martinet, André (1964). *Économie des changements phonétiques; Traité de phonologie diachronique*. 2nd ed. Berne: Francke.

Neef, Martin and Miriam Balestra (2011). "Measuring graphematic transparency. German and Italian compared." In: *Written Language and Literacy* I.14, pp. 109–142.

Oliveira, Daniela Santos et al. (2016). "Effects of language experience on the discrimination of the Portuguese palatal lateral by nonnative listeners." In: *Clinical Linguistics & Phonetics* 30.8, pp. 569–583.

Paap, Kenneth R. and Ronald W. Noel (1991). "Dual-route models of print to sound: Still a good horse race." In: *Psychological Research* 53.1, pp. 13–24.

Presutti, Stefano (2019). "L'interdépendance entre oralité et écriture: le cas de la consonne latérale palatale dans l'acquisition phonologique de l'italien langue étrangère." PhD thesis. Université Aix-Marseille.

Rastle, Kathleen and Max Coltheart (1998). "Whammies and double whammies: The effect of length on nonword reading." In: *Psychonomic Bulletin & Review* 5.2, pp. 277–282.

Richlan, Fabio (2014). "Functional neuroanatomy of developmental dyslexia: the role of orthographic depth." In: *Frontiers in Human Neuroscience* 8.

Schmalz, Xenia et al. (2015). "Getting to the bottom of orthographic depth." In: *Psychonomic Bulletin & Review* 22.6, pp. 1614–1629.

Seidenberg, Mark S. (1992). "Beyond Orthographic Depth in Reading: Equitable Division of Labor." In: *Advances in Psychology*. Ed. by Ram Frost and Leonard Katz. Vol. 94. Orthography, Phonology, Morphology, and Meaning. North-Holland, pp. 85–118.

Share, David L. (2008). "On the Anglocentricities of current reading research and practice: the perils of overreliance on an "outlier" orthography." In: *Psychological Bulletin* 134.4, pp. 584–615.

Stevens, Kenneth N. (1989). "On the quantal nature of speech." In: *Journal of Phonetics* 17, pp. 3–46.

Tresoldi, Martina et al. (2018). "Normative and validation data of an articulation test for Italian-speaking children." In: *International Journal of Pediatric Otorhinolaryngology* 110, pp. 81–86.

Tressoldi, Patrizio E., Cesare Cornoldi, and Anna M. Re (2017). *BVSCO-2: batteria per la valutazione della scrittura e della competenza ortografica-2: manuale e materiali per le prove*. Florence: Giunti Edu.

Trubetzkoy, Nicolai (1971). *Principles of phonology*. Berkeley: University of California Press.

Ziegler, Johannes C. (2018). "Différences inter-linguistiques dans l'apprentissage de la lecture." In: *Langue française* 3, pp. 35–49.

Ziegler, Johannes C., Conrad Perry, and Max Coltheart (2000). "The DRC model of visual word recognition and reading aloud: An extension to German." In: *European Journal of Cognitive Psychology* 12.3, pp. 413–430.

A Small Step for a Man,
a Giant Leap for a People
—The Coptic Alphabets

Victoria Fendel

Abstract. The paper looks at the beginnings of the Coptic alphabet in first- and second-century Egypt from different angles. It reviews and builds on the sometimes-contradictory research from the social perspective while also considering practical challenges for the ancient writers. It explores the relevance of cognitive factors regarding the transition to the first alphabetic writing system for the Egyptian language.

In the later Roman period (1st / 2nd c. AD), new writing systems to notate the Egyptian language emerged, which were to suit the needs of the evolved Egyptian language. The Greek alphabet and the Demotic script served as their models and resources. The impetus for this change must be sought in the social setting. The paper pulls together sociolinguistic and empirical past research and adds the cognitive-linguistic angle. This contribution is not a complete account of the argument, but a deeper dive into three issues that sharpen aspects of the argument made and the hypothesis put forward.

Egyptian could be written with three writing systems for most of its history. These were a cursive for day-to-day writing, a script primarily used in religious contexts, and a script used for monumental inscriptions (Houston, Baines, and Cooper, 2003, pp. 440–442). The day-to-day writing system has traditionally lent its name to the stage of Egyptian since Ptolemaic times. Thus, we speak of Demotic in the Ptolemaic and early Roman periods and of Coptic in the later Roman and early Byzantine periods. These labels are solely owing to research traditions; the Egyptian language developed continuously.

Egypt came under Ptolemaic rule in the aftermath of Alexander's victories (4th c. BC) and under Roman rule following Octavian's / Augustus' victory at Actium (1st c. BC). The later Roman period saw a political, societal and cultural turnover. Politically, the central power weakened, as

Victoria Fendel   ⓘD 0000-0001-6302-3726
University of Oxford (Leverhulme Early Career Fellow), Lady Margaret Hall, Norham Gardens, Oxford OX2 6QA, United Kingdom
E-mail: victoria.fendel@classics.ox.ac.uk

evident in the requirements for valid wills after the Constitutio Antonini-
ana (AD 212) granted citizenship to many inhabitants of the empire. The
requirements related to the distinction between Roman and non-Roman
wills disappeared and Egyptian became a permissible language (Garel
and Nowak, 2017). Societally, we see mixture. Naming practices regard-
ing double names indicate a thorough mixture of the Greek and Egypt-
ian social brackets in everyday life. Successful individuals such as the no-
tary Hermias (Vierros, 2012) in Ptolemaic Pathyris, and the businessman
Phoibammon in early Byzantine Aphrodito (Keenan, 2007, pp. 233–237)
confirm Kraus' (2000) hypothesis that social brackets were increasingly
determined by wealth rather than ethnicity. Economically, we observe
decline. The building activity in villages and cities decreased and the
defence system finally crumbled under the Sassanid attacks (7th c. AD)
(Foss, 2003; Keenan, 2007; Kiss, 2007; Sänger, 2011; van Minnen, 2007).
In this setting, local and clerical, Christian, institutions took on new tasks
in the educational and administrative spheres (Fournet, 2019, Chapter 4;
Quack, 2017b; Wipszycka, 2007).

The Coptic alphabet emerged as the most salient strand amongst sev-
eral writing systems, which drew on the Greek alphabet and the Demotic
script as models and graphemic resources (Quack, 2017a). Apparently,
a range of small communities of practice[1] developed writing systems in
this period of time, potentially motivated by the general atmosphere of
change. The predominance of the Coptic alphabet results from social
factors, whereas the origins of the Coptic alphabets are also linked to
the cognitive concept of a best fit between a writing system and the lan-
guage to be represented.[2]

Socio-linguistically, we notice the imbalance between Greek and
Egyptian with regard to Matras' (2009) criteria for a language to be
successful in a bilingual setting, which we extend to a writing system:
a writing system, educational backing, and political backing. Quack
(2017a) and Choat (2012) prove wrong convincingly Bagnall's (1993) hy-
pothesis that there was a gap of about 150 years between the disappear-
ance of the Demotic and the emergence of the Coptic scripts. Nonethe-
less, at the time, the Greek alphabet was an established writing system
with educational and political backing, whereas the Coptic script was in
its infancy.

---

1. A community of practice is a group of people who is engaging in exchange of
knowledge and practices. Prime examples are schools (Unwin, Hughes, and Jewson,
2007).

2. Alternative hypotheses: Demotic was no longer fit for the current stage of the
language (Dieleman, 2005, p. 71; Stadler, 2008, pp. 159–160), Coptic evolved as an in-
group writing system (Bagnall, 2005; Choat, 2012, p. 588; Quack, 2017a, p. 73; Torallas
Tovar, 2004b, p. 59; Torallas Tovar and Vierros, 2019, p. 488), Coptic evolved in the
context of local nationalistic uprisings (Choat, 2009, p. 354; Clackson, 2010, p. 94).
As the references show, all of these have been refuted.

Socio-culturally, we see, as mentioned, contact in day-to-day life intensify. This is not least evident in the widespread use of loanwords, from Egyptian into Greek and vice versa, referring to everyday realities (Förster, 2002; Torallas Tovar, 2004a,b; 2007; 2017).[3] The use of loanwords is particularly common in the context of Christianity, the rise of which constitutes the most fundamental cultural change of the time (Edict of Milan, AD 313, Edict of Thessalonika, AD 380) (Depauw and Clarysse, 2013; Houston, Baines, and Cooper, 2003). In examples such as 'father', the Egyptian word remained the everyday word, whereas the Greek loanword acquired a specialised Christian meaning.

These sociolinguistic and sociocultural settings make drawing upon the Greek alphabet as a model and resource comprehensible. Yet, they also show how in-groups, such as the early Christians, could promote new writing systems very successfully. By contrast, the origins of the systemic change from a largely supraphonemic to a phonemic writing system is linked to phonological changes affecting the fit of the writing system to the language, conceptualised in the grain size theory. We call supraphonemic a writing system that maps phonological units such as syllables onto graphemes; we call phonemic a writing system that maps phonemes onto graphemes (Perfetti and Verhoeven, 2017, p. 23). The inherited Egyptian scripts were mixed. They combined consonantal, biconsonantal and triconsonantal phonograms, ideogrammatic logograms and determinatives (Gardiner, 1957, paras 6, 17, 22, 23).[4] Some vowels could be indicated (i.e., ꜣ ≈ /a/, j ≈ /i/, and w ≈ /u/), but vowel writing was not consistent.

Classical Egyptian 𓇋𓏏𓀀 *jt*
Demotic 𓊪 *it*
Coptic ⲈⲒⲰⲦ (S) *eiōt* / ⲒⲰⲦ (B) *iōt*

FIGURE 1. Non-alphabetic vs. alphabetic writing systems

The absence of a one-to-one correspondence between a script and a writing system makes possible changes in the writing system for any language. Any writing system is a reduction of the acoustic signal it represents based on the principles of economy and practicability. However, in theory, there is a best fit between a writing system and a language

---

3. There is no borrowing of inflectional morphology, which would point to renegotiating of identity (Matras, 2015).

4. Some signs can be phonograms or logograms, that is indicate a sound or a meaning. Determinatives are added to a phonologically represented word in order to disambiguate the meaning. For example, *nb* ◡ 'everyone' vs. ◡𓀭 'lord' (Ockinga, 2012, p. 2).

based on the size of the unit mapped onto a grapheme, that is the grain size (Baroni, 2011). In reality, this ideal fit might once have existed in the history of a writing system with a language but disappears when the language develops phonologically but the writing system is attached to it for non-linguistic reasons (cf. tradition, etc.).

The eventual Coptic alphabet is an elaborate adaptation of its Greek model. It is based on the Koine Greek alphabet rather than the earlier local Greek alphabets judging by the sound-grapheme mappings and inventory of graphemes (Horrocks, 2014, p. 170; Jeffery, 1990). The initial development was decentralised as not only Quack's (2017) observations regarding the regionalisation of Demotic in the preceding period, but primarily the letter shapes and inventories of Demotic-derived signs show.[5] Quack (2017a) assumes that a functioning version of the Coptic alphabet was in circulation by AD 100, and the relaxation of linguistic norms in AD 212 would have helped promotion of any new script in circulation.

ⲁ	ⲃ	ⲅ	ⲇ	ⲉ	ϛ	ⲍ	ⲏ
/a/	/b/ (S) /v/ (B, A)	/g/	/d/	/e/		/z/	/ē/

ⲑ	ⲓ	ⲕ	ⲗ	ⲙ	ⲛ	ⲝ	ⲟ
/t/+/h/	/y/	/k/	/l/	/m/	/n/	/k/+/s/	/o/

ⲡ	ⲣ	ⲥ	ⲧ	ⲩ / ⲟⲩ	ⲫ	ⲭ	ⲯ
/p/	/r/	/s/	/t/	/w/	/p/+/h/	/k/+/h/	/p/+/s/

ⲱ	ⲟⲩ	ϥ	ⳉ (B) /ϧ (A)	ϩ	ϫ	ϭ	ϯ
/ō/	/š/	/f/	/x/	/h/	/č/	/k^y/	/t/+/y/

FIGURE 2. The Coptic alphabet (cf. Layton, 2011, paras 8, 13)

Here, we turn to the three deeper-dive issues concerning the societal, phonological and cognitive aspects of the argument.

1. Society—Literacy Rates:
Were Literacy Rates Favouring the Greek Alphabet?

The short answer to this question is yes, literacy rates were most probably favouring the Greek alphabet. The Greek alphabet was an estab-

---

5. Demotic-derived signs are those that are based on Demotic signs but likened to the other alphabetic signs, for instance with regard to filling roughly a rectangle on the line (Quack, 2017a).

lished writing system in the Roman period. There was full educational and political backing for it. By contrast, Egyptian writing was on the way back up. Educational facilities were in the making (cf. monasteries); educational centres had to be established for the new writing system as the old temples, which were the educational centres for Demotic, were losing funding and status (Cribiore, 2001; Houston, Baines, and Cooper, 2003; Maehler, 1983). Different writing systems were still competing, and thus none of them had yet attained the status of a standard writing system. Political backing was still lacking (Fournet, 2019).[6] Furthermore, the contexts for use of writing were more extensive for Greek than for Egyptian due to the administrative apparatus. Depauw (2009; 2012) has described extensively how Greek had taken over from Demotic.

Given the educational situation as well as the predominance of Greek in administrative circles, it is likely that literacy rates in Greek were significantly higher than literacy rates in Egyptian. This is where Stadler's (2008, pp. 166–167) critical mass argument comes in. According to this, a writing system needs to be used by a significant number of people not only to stay alive as it were but also in order to be useful—sender and addressee need to be able to operate in the same writing system. This situation may have favoured the Greek alphabet.

2.   Phonology—Vowel Writing: Was There Pressure to Start Writing Vowels?

Overall, the impression is that there was some pressure to start writing vowels.

Firstly, earlier Egyptian already notated vowels occasionally in the form of the *matres lectionis*. *Matres lectionis* are signs that indicate a vowel in writing systems that do not notate vowels consistently. The relevant signs in Egyptian are aleph, iod and waw. They can represent a consonant or a vowel, but as *matres lectionis* always indicate vowels (Hornkohl and Khan, 2020; Werning, 2016).

Secondly, several systems experimenting with vowel writing in Egyptian competed in the early Roman period. Quack (2017a) lists (i) the Greek alphabet / Graeco-Egyptian, (ii) Demotic syllabic signs / syllabic writing, (iii) the Greek alphabet with some Demotic signs, (iv) Demotic mono-consonantal signs / alphabetic Demotic, and (v) the Greek alphabet with Demotic-derived signs / Old Coptic.

Thirdly, changes in the Egyptian syllable structure, such as an increase in open syllables and the development of biconsonantal onsets

---

6. Political backing refers to the acceptability of a language and writing system in all registers including highly formal ones.

(Allen, 2013, pp. 13, 24; Loprieno, 1995, pp. 36–37), had made it increasingly difficult to use a supraphonemic writing system. The Universal Phonological Principle states that phonological information is accessed before lexical information when reading (Baroni, 2011; Gleitman, 1985), including in shallow orthographies such as Hebrew (Frost, 1994). Thus, we prefer a writing system that represents the phonology of a language at least approximately.

Finally, practically speaking, Greek loanwords were frequent in everyday and Christian contexts. They were difficult to transcribe into Demotic as their small number proves (Clarysse, 1987; 2013; Ray, 1994). In essence, one had to delete the vowels while ensuring that the string remained a unique signifier of the meaning and choose a determinative (Crellin, 2018). This seems disadvantageous in a thoroughly bilingual environment.

ἀποχή *apok^b ē* 'receipt'
(a) 3pwg^c [bookroll determinative] (P. Berl. 8043 verso 3.20; 4.10)
(b) pg^c [bookroll determinative] (JEA 55, 1969, 187)

Figure 3. Loanwords (cf. Clarysse, 2013)

Overall, there is no complete change of systems, but a move from some vowel writing to consistent vowel writing. Competing systems evolved around the same idea. The changing political and societal settings may have offered opportunities for smaller groups to experiment with an until then traditional 'untouchable' writing system. These same political and societal settings allowed the Coptic alphabet to win out eventually.

3.   Cognition—Best Fit: Is One Script More Suitable for Representing a Language Than Another?

According to the grain size model, there is a better (if not a best) fit between a language and a writing system. The grain size of a writing system is determined by (i) pressures towards smaller and orthographically less complex units (i.e., granularity), (ii) pressures towards larger and phonologically more accessible units (i.e., availability), and (iii) pressures towards maximally consistent units (i.e., consistency) (Asfaha, Kurvers, and Kroon, 2009; Ziegler and Goswami, 2005). The phonolog-

ical structure of a language will favour one or the other type of writing system.[7]

However, there are pressures towards hanging on to a writing system even if the fit between language and writing system is not perfect. Firstly, users are familiar with the mapping principles of their writing system (Perfetti and Dunlap, 2008) and have to acquire new mapping principles when learning a new writing system (Bassetti, 2016; Hirshorn and Fiez, 2014; Keiko, 2002; Lallier and Carreiras, 2018). Secondly, cultural, social and political pressures impact on updating vs. preserving and switching vs. retaining a writing system. Some relevant aspects include the readability by regular interlocutors[8], access to training, the prestige and cultural significance attached to a script[9], and the resources using this script that would have to be modified.[10] In fact, Thomason's (2001) argument that attitudinal factors override linguistic factors with regard to language change could be transferred onto script change. Her claim has been variously contested, yet attitudinal factors are far from irrelevant.

A prime example of a writing system people hung on to is Demotic, which is often described as conservative (Depauw, 1997, p. 36; Oréal, 1999, p. 295; Richter, 2009, p. 403; Thompson, 2009, p. 399), yet remained the Egyptian writing system until Coptic emerged.

References

Allen, J. (2013). *The ancient Egyptian language: An historical study*. Cambridge: Cambridge University Press.

Asfaha, Y., J. Kurvers, and S. Kroon (2009). "Grain Size in Script and Teaching: Literacy Acquisition in Ge'ez and Latin." In: *Applied Psycholinguistics* 30, pp. 709–734.

Bagnall, R. (1993). *Egypt in late antiquity*. Princeton: Princeton University Press.

———— (2005). "Linguistic Change and Religious Change: Thinking about the Temples of the Fayoum in the Roman Period." In: *Christianity and Monasticism in the Fayoum Oasis: Essays from the 2004 International Symposium of the Saint Mark Foundation and the Saint Shenouda the Archimandrite Coptic Society in Honor of Martin Krause*. Ed. by G. Gabra. Cairo: American University in Cairo Press, pp. 11–19.

---

7. Perfetti and Verhoeven (2017, p. 23) list syllabic, morpho-syllabic, alphasyllabic, abjad and alphabetic writing systems.

8. A modern example is transcriptions of Chinese (Chappell, 1980).

9. A modern example might be the Greek alphabet in modern-day Europe.

10. A literary example is Orwell's 1984.

Baroni, A. (2011). "Alphabetic vs. Non-alphabetic writing: Linguistic fit and natural tendencies." In: *Rivista Di Linguistica* 23.2, pp. 127–159.

Bassetti, B. (2016). "Learning second language writing systems." LLAS—Centre for Languages, Linguistics and Area Studies `https://www.llas.ac.uk/resources/gpg/2662.html`.

Chappell, H. (1980). "The Romanization Debate." In: *The Australian Journal of Chinese Affairs* 4, pp. 105–118.

Choat, M. (2009). "Language and Culture in Late Antique Egypt." In: *A Companion to Late Antiquity*. Ed. by P. Rousseau and J. Raithel. Chicester: John Wiley & Sons, pp. 342–356.

———— (2012). "Coptic." In: *The Oxford handbook of Roman Egypt*. Ed. by C. Riggs. Oxford: Oxford University Press, pp. 581–593.

Clackson, S. (2010). "Coptic or Greek? Bilingualism in the papyri." In: *The multilingual experience in Egypt: From the Ptolemies to the cAbbasids*. Ed. by A. Papaconstantinou. Farnham, Surry, UK: Ashgate Publishing, pp. 73–104.

Clarysse, W. (1987). "Greek loan-words in Demotic." In: *Aspects of Demotic lexicography: Acts of the second international conference for Demotic studies*. Ed. by S. Vleeming. Leiden: Peeters, pp. 9–33.

———— (2013). "Determinatives in Greek loan-words and proper names." In: *Aspect of Demotic orthography; Acts of an International colloquium held in Trier*. Ed. by S. Vleeming. Peeters. Leiden, pp. 1–24.

Crellin, R. (2018). "Measuring ambiguity and the invention of vowel-writing in Greek." In: *Proceedings of International Colloquium of Ancient Greek Linguistics ICAGL 9*. Ed. by M. Leiwo, M. Vierros, and H. Halla-aho. Helsinki.

Cribiore, R. (2001). *Gymnastics of the mind: Greek education in Hellenistic and Roman Egypt*. Princeton: Princeton University Press.

Depauw, M. (1997). *A companion to demotic studies*. Vol. 28. Papyrologica Bruxellensia. Brussels: Fondation égyptologique reine Élisabeth.

———— (2009). "Bilingual Greek-Demotic Documentary Papyri and Hellenization in Ptolemaic Egypt." In: *Faces of Hellenism. Studies in the History of the Eastern Mediterranean (4th century BC-5th century AD)*. Ed. by P. Van Nuffelen. Vol. 48. Studia Hellenistica. Leiden: Peeters, pp. 120–139.

———— (2012). "Language use, literacy and bilingualism." In: *The Oxford Handbook of Roman Egypt*. Ed. by C. Riggs. Oxford: Oxford University Press, pp. 493–506.

Depauw, M. and W. Clarysse (2013). "How Christian was Fourth Century Egypt? Onomastic Perspectives on Conversion." In: *Vigiliae Christianae* 67.4, pp. 407–435.

Dieleman, J. (2005). *Priests, tongues, and rites: The London-Leiden magical manuscripts and translation in Egyptian ritual (100–300 CE)*. Vol. 153. Religions in the Graeco-Roman World. Leiden, Boston: Brill.

Förster, H. (2002). *Wörterbuch der griechischen Wörter in den koptischen dokumentarischen Texten*. Vol. 148. Texte und Untersuchungen zur Geschichte der altchristlichen Literatur. Berlin: Mouton De Gruyter.

Foss, C. (2003). "The Persians in the Roman near East (602–630 AD)." In: *Journal of the Royal Asiatic Society* 13.2, pp. 149–170.

Fournet, J.-L. (2019). *The rise of Coptic: Egyptian versus Greek in late antiquity*. Princeton: Princeton University Press.

Frost, R. (1994). "Prelexical and postlexical strategies in reading: Evidence from a deep and a shallow orthography." In: *Journal of Experimental Psychology. Learning, Memory, and Cognition* 20.1, pp. 116–129.

Gardiner, A. (1957). *Egyptian grammar: Being an introduction to the study of hieroglyphs*. 3rd ed. Oxford: Oxford University Press.

Garel, E. and M. Nowak (2017). "Monastic Wills: The Continuation of Late Roman Legal Tradition?" In: *Writing and Communication in Early Egyptian Monasticism*. Ed. by M. Choat and M. Giorda. Brill. Leiden, Boston, pp. 108–128.

Gleitman, L. (1985). "Orthographic Resources Affect Reading Acquisition—If They Are Used." In: *Remedial and Special Education* 6.6, pp. 24–36.

Hirshorn, E. and J. Fiez (2014). "Using artificial orthographies for studying cross-linguistic differences in the cognitive and neural profiles of reading." In: *Journal of Neurolinguistics* 31, pp. 69–85.

Hornkohl, A. and G. Khan (2020). *Studies in Semitic vocalisation and reading traditions*. Cambridge: Open Book Publishers.

Horrocks, G. (2014). *Greek: A history of the language and its speakers*. 2nd ed. Chichester: Wiley Blackwell.

Houston, S., J. Baines, and J. Cooper (2003). "Last Writing: Script Obsolescence in Egypt, Mesopotamia, and Mesoamerica." In: *Comparative Studies in Society and History* 45.3, pp. 430–479.

Jeffery, L. (1990). *The local scripts of archaic Greece: A study of the origin of the Greek alphabet and its development from the eighth to the fifth centuries BC*. Oxford: Clarendon Press.

Keenan, J. (2007). "Byzantine Egyptian villages." In: *Egypt in the Byzantine world*. Ed. by R. Bagnall. Cambridge: Cambridge University Press, pp. 300–700.

Keiko, K. (2002). "Writing Systems and Learning to Read in a Second Language." In: *Chinese Children's Reading Acquisition: Theoretical and Pedagogical Issues*. Ed. by L. Wenling, J. Gaffney, and J. Packard. Boston, MA: Springer, pp. 225–248.

Kiss, Z. (2007). "Alexandria in the fourth to seventh centuries." In: *Egypt in the Byzantine world*. Ed. by R. Bagnall. Cambridge: Cambridge University Press, pp. 300–700.

Kraus, T. (2000). "(Il)literacy in non-literary papyri from Graeco-Roman Egypt: Further aspects of the educational ideal in ancient literary sources and modern times." In: *Mnemosyne* 53.3, pp. 322–342.

Lallier, M. and M. Carreiras (2018). "Cross-linguistic transfer in bilinguals reading in two alphabetic orthographies: The grain size accommodation hypothesis." In: *Psychon Bulletin Review* 25, pp. 386–401.

Layton, B. (2011). *A Coptic Grammar*. Porta Linguarum Orientalium. Wiesbaden: Harrassowitz.

Loprieno, A. (1995). *Ancient Egyptian: A linguistic introduction*. Cambridge: Cambridge University Press.

Maehler, H. (1983). "Die griechische Schule im ptolemäischen Ägypten." In: *Egypt and the Hellenistic world: Proceedings of the international colloquium Leuven—24–26 May 1982*. Ed. by E. van't Dack, P. van Dessel, and W. van Gucht. Vol. 27. Studia Hellenistica. Leuven: Peeters, pp. 191–203.

Matras, Y. (2009). *Language contact*. Cambridge: Cambridge University Press.

———— (2015). "Why is the borrowing of inflectional morphology dispreferred?" In: *Borrowed Morphology*. Ed. by F. Gardani, P. Arkadiev, and N. Amiridze. Berlin: Mouton de Gruyter.

Ockinga, B. (2012). *A concise grammar of Middle Egyptian: An outline of Middle Egyptian grammar*. 3rd ed. Mainz, Germany: Philipp Von Zabern.

Oréal, E. (1999). "Contact linguistique. Le cas du rapport entre le grec et le copte." In: *Lalies* 19, pp. 289–306.

Perfetti, C. and S. Dunlap (2008). "Learning to read: General principles and writing system variations." In: *Learning to Read Across Languages: Cross-Linguistic Relationships in Firstand Second-Language Literacy Development*. Ed. by K. Koda and A. Zehler. Abington-on-Thames: Routledge, pp. 13–38.

Perfetti, C. and L. Verhoeven (2017). "Introduction: Operating Principles in Learning to Read." In: *Learning to Read across Languages and Writing Systems*. Ed. by C. Perfetti and L. Verhoeven. Cambridge: Cambridge University Press, pp. 1–30.

Quack, J. (2017a). "How the Coptic script came about." In: *Greek influence on Egyptian-Coptic*. Ed. by E. Grossman et al. Vol. 17. Lingua Aegyptia. Hamburg: Widmaier, pp. 27–96.

———— (2017b). "On the Regionalization of Roman-Period Egyptian Hands." In: *Scribal Repertoires in Egypt from the New Kingdom to the Early Islamic Period*. Ed. by J. Cromwell and E. Grossman. Oxford: Oxford University Press, pp. 184–211.

Ray, J. (1994). "How demotic is Demotic?" In: *Egitto e Vicino Oriente* 17, pp. 251–264.

Richter, T. (2009). "Greek, Coptic and the 'language of the Hijra': The rise and decline of the Coptic language in late antique and medieval Egypt." In: *From Hellenism to Islam: Cultural and Linguistic Change in the Roman Near East*. Ed. by H. Cotton. Cambridge: Cambridge University Press, pp. 401–446.

Sänger, P. (2011). "The Administration of Sasanian Egypt: New Masters and Byzantine Continuity." In: *Greek, Roman, and Byzantine Studies* 51.4, pp. 653–665.

Stadler, M. (2008). "On the Demise of Egyptian Writing. Working on a Problematic Source Basis." In: *The Disappearance of Writing Systems: Perspectives on literacy and communication*. Ed. by J. Baines, J. Bennett, and S. Houston. Oxford: Equinox, pp. 157–181.

Thomason, S. (2001). *Language contact*. Edinburgh: Edinburgh University Press.

Thompson, D. (2009). "The multilingual environment of Persian and Ptolemaic Egypt: Egyptian, Aramaic and Greek documentation." In: *The Oxford Handbook of Papyrology*. Ed. by R. Bagnall. Oxford: Oxford University Press, pp. 395–417.

Torallas Tovar, S. (2004a). "Egyptian lexical interference in the Greek of Byzantine and early Islamic Egypt." In: *Papyrology and the history of early Islamic Egypt*. Ed. by P. Sijpesteijn and L. Sundelin. Vol. 55. Islamic history and civilization. Leiden: Brill, pp. 163–198.

————— (2004b). "The context of loanwords in Egyptian Greek." In: *Lenguas en contacto: El testimonio escrito*. Ed. by P. Bádenas de la Pena et al. Vol. 46. Manuales y anejos de "Emerita". Madrid: Consejo Superior de Investigaciones Científicas, pp. 57–67.

————— (2007). "Egyptian loanwords in Septuaginta and the papyri." In: *Akten des 23. Internationalen Papyrologenkongresses, Wien*. Ed. by B. Palme. Wien: Verlag der Österreichischen Akademie der Wissenschaften, pp. 687–692.

————— (2017). "The Reverse Case: Egyptian Borrowing in Greek." In: *Greek Influence on Egyptian Coptic: Contact induced change in an ancient African language*. Ed. by E. Grossman et al. Vol. 17. Lingua Aegyptia. Hamburg: Widmaier, pp. 97–113.

Torallas Tovar, S. and M. Vierros (2019). "Languages, Scripts, Literature, and Bridges Between Cultures." In: *A Companion to Graeco-Roman and Late Antique Egypt*. Ed. by K. Vandorpe. Chichester: Wiley Blackwell, pp. 483–499.

Unwin, L., J. Hughes, and N. Jewson (2007). *Communities of practice: Critical perspectives*. London: Routledge.

van Minnen, P. (2007). "The other cities in Later Roman Egypt." In: *Egypt in the Byzantine world*. Ed. by R. Bagnall. Cambridge: Cambridge University Press, pp. 300–700.

Vierros, M. (2012). *Bilingual notaries in Hellenistic Egypt: A study of Greek as a second language*. Vol. 5. Collectanea Hellenistica. Leiden: Peeters.

Werning, D. (2016). "Hypotheses on glides and matres lectionis in earlier Egyptian orthographies." In: *Coping with obscurity: The Brown workshop on earlier Egyptian grammar*. Ed. by J. Allen, M. Collier, and A. Stauder. Vol. 4. Wilbour Studies in Egyptology and Assyriology. Atlanta, GA: Lockwood Press, pp. 29–44.

Wipszycka, E. (2007). "The institutional church." In: *Egypt in the Byzantine world*. Ed. by R. Bagnall. Oxford: Oxford University Press, pp. 300–700.

Ziegler, J. and U. Goswami (2005). "Reading Acquisition, Developmental Dyslexia, and Skilled Reading Across Languages: A Psycholinguistic Grain Size Theory." In: *Psychological Bulletin* 131.1, pp. 3–29.

Old Aramaic Script in Georgia

Helen Giunashvili

Abstract. Old Aramaic and its script are most important to the history of the Georgian culture. On the territory of contemporary Georgia, particularly in its Eastern part, being historically Iberian kingdom (4th c. BC−4th c. AD), a number of original Aramaic inscriptions are found. They are inscribed on different objects and could be dated by the period of 3rd c. BC−3rd c. AD.

The greater part of these Aramaic inscriptions is executed in a variety of the North-Mesopotamian type of Aramaic script, known as "Armazian," one of the outgrowths of the Imperial (Official) Aramaic writing, widely used in Achaemenid Empire (550 BC−330 BC).

The whole corpus of the Aramaic inscriptions of Georgia requires systematic interdisciplinary researches, for revealing the main trends of its typological development in the light of Near Eastern-South Caucasian cultural-linguistic interference.

1.   Introduction

Aramaic is of great importance for Georgia. All the three historical phases of this language: Old, Middle and Modern are well represented in the Georgian cultural tradition (K. Tsereteli, 1994).

On the territory of contemporary Georgia, mainly in its Eastern part−Kartli historically Iberian kingdom (4th c. BC−4th c. AD), a number of original Aramaic inscriptions were found.

They were made on different objects: steles (an epitaph and a victory stele), bone plates, wine-pitchers, silver bowls, and household items, stones of sanctuary buildings and sarcophagi, jewels. For the present, the whole corpus of inscriptions comprises nearly 100 units dated by 3rd−2nd c. BC−3rd c. AD (K. Tsereteli, 1998b) and kept at different

Helen Giunashvili   (iD) 0000-0003-1271-8354
G. Tsereteli Institute of Oriental Studies/Ilia State University, 3 George Tsereteli St.,
Tbilisi 0162, Georgia
E-mail: elene.giunashvili@iliauni.edu.ge

funds of the National Museum of Georgia.[1] These ancient Aramaic inscriptions were discovered in Mtskheta, the capital of Iberia, as well as its outskirts—Armazi, Bagineti and other various locations in Central Georgia—Uplistsikhe, Urbnisi, Zghuderi, Bori, Dedoplis Gora (Mindori), Dzalisa.[2]

The Aramaic inscriptions of Georgia are distinguished by their form and content. Some of them are quite extensive, such as Armazi steles and a number of dedicatory inscriptions dated by 3rd c. AD, found on golden bracelets from Armazi burials. The rest of inscriptions are rather short, consisting only of one or two words, denoting a proper name, or a title, they also frequently have an attributive meaning of weight, size and function of an object. The great part of these inscriptions still remains unpublished (K. Tsereteli, 1998b).

The Old Aramaic was one of chief written languages of Iberia before the adoption of Christianity (4th century AD).

Later Aramaic epigraphic monuments (4th–5th c. AD), also revealed in Mtskheta, belong to a particular category. They were created by the Jewish community of Mtskheta, and are written in Hebrew characters, while their language is Aramaic (Jewish-Palestinian dialect) (G. Tsereteli, 1962, pp. 377–378; K. Tsereteli, 1996; 1998a; Shaked, 2006).

Origins of spreading the Aramaic language in Georgia and, generally, in the South Caucasus are to be traced in the Achaemenian epoch (6th–4th c. BC) of the Persian Empire, when firm foundations of Iranian statehood and national culture were laid, and it was widely used as the official language of the Empire.

The most ancient Aramaic inscription found in Georgia, is the inscription on the silver bowl from the Kazbegi treasure (5th c. BC), being a specimen of Achaemenid art, brought into this region of the Iranian dominance. Most scholars considered the bowl's inscription as a proper name of its owner (K. Tsereteli, 2001b) (Fig. 1).[3]

Iberian kingdom and Kartvelian tribes are not mentioned in the extant Old Persian inscriptions; however, rich historical-archaeological material and linguistic-philological evidences testify the strong Iranian cultural impact on this region.

---

1. For the first complete list of these inscriptions see Gagoshidze and Tsotselia, 1991, pp. 71–72; cf. also Giorgadze, 2008, pp. 253–255.

2. All these geographic sites are important historical places, where the most valuable archaeological discoveries have been made. On their description and catalogue, see Furtwängler, Gagoshidze, Löhr, and Ludwig, 2008, pp. 257–272.

3. The bowl was found in Georgia, in the village of Stepantsminda (Kazbegi) and is kept at present at the State Historical Museum of Russia (inventory number SB1735, weight $\sim$ 266,5 grams). The Aramaic inscription and the photo of this bowl was first included in the Corpus Inscriptionum Semiticarum, 1889, and later in the Corpus of Canaanite and Aramaic Inscriptions by Donner and Röllig, 1969.

Figure 1. Kazbegi silver bowl with the Aramaic inscription, photos reprinted from Smirnov (1909, N 13, Table III).

The introduction of administrative, social, political and legal institutions evolved in the Achaemenid Empire in the South Caucasus was of great significance. These institutions and socio-economic processes taking place in the Achaemenid period played an important role in the emergence and development of the Iberian and Armenian kingdoms (G. Tsereteli, 1974).

Medieval Georgian chronicles (11th c.) preserve particularly valuable data on this subject. One of them, *The Life of Kartli* (consisting of multiple sources several of which are of remarkable antiquity) narrates that the first Georgian king Parnavaz (Pharnabazos, Greek Φαρνάβαζος),[4] who was a representative of a powerful aristocratic family from Mtskheta and was coroneted about 280 BC., created his state "like the kingdom of the Persians" (Qaukhchishvili, 1955, p. 21; Metreveli, 2008, p. 44).

One of the chapters of *The Life of Kartli* dealing with the life and deeds of the Georgian king mentions Aramaic among languages widespread in pre-Christian Iberia: "Six languages were spoken in Kartli: *Armenian*, *Georgian*, *Khazar*, *Assyrian* (i.e., Aramaic), *Hebrew* and *Greek*. And all the kings of Kartli and all the men and women, knew these languages" (Qaukhchishvili, 1955, p. 16; Metreveli, 2008, p. 36).[5]

———————————

4. On the Iranian etymology of this royal name in Georgian, see Andronikashvili, 1966, pp. 496–499; Chkheidze, 1984, pp. 32–33, 47. For the complete bibliography on this name, etymological studies and its penetration in Armenian, see Martirosyan, to appear.

5. On the use of the term Assyrian (language) in the meaning of Aramaic in Old Georgian, see K. Tsereteli, 1976, pp. 184–185.

(a)   (b)   (c)

FIGURE 2. (a) A portrait of King Parnavaz done by the Georgian painter Zurab Kapanadze (1924–1989). (b) Armazi Monolingual, photo made at the Stone Fund of the Georgian National Museum. (c) Armazi Bilingual, photo made at the Stone Fund of the Georgian National Museum.

Georgian historiography ascribes to the first legendary king Parnavaz the creation of the Georgian writing ("Georgian literacy") (Fig. 2a).

According to Professor Thomas Gamkrelidze's theory (Gamkrelidze, 1989), the "Georgian literacy" might have meant its introduction in the form of the so-called "alloglottography" or "writing-in-another-language" widely used in the Achaemenian chancelleries[6], i.e., reading a text written in some widespread foreign language, in this case Aramaic, on the basis of the local language (the Georgian), before introducing the national script.

The existence of "literary traditions" in the pre-Christian Georgian World, where Old Aramaic alongside with Greek were widely used, should be assumed in the form of oral tradition and folklore. The introduction of national writing when Christianity was proclaimed as the state religion only served to record such tradition, and further strengthen and develop the literary language.[7]

---

6. The term "alloglottography" was established in *Ancient Iranian Studies* by Ilya Gershevitch (1979, pp. 114–119). For the modern studies of alloglottography in the Ancient Near Eastern cultural tradition, see Rubio (2006).

7. Such oral traditions were strengthened also by a rule of rendering the Scripture in the newly Christianized Eastern world (and probably in Georgia, too) that may be called "Alloglottoepy" or "Saying-in-another-language," when religious texts (of the Old, and especially New, Testament) were preached directly through oral rendering and translation. This contributed to the refinement and enrichment of the vocabu-

The most ancient Georgian literary monuments are dated only by the 5th c. AD, the period when the written translation of the Scripture into Georgian has already been realized and recorded in the Old Georgian original script, Asomtavruli.

The Aramaic script used in Iberia passed a long way of development. It was one of the outgrowths of the Imperial (Official) Aramaic writing, widely used in Achaemenid Empire (550–330 BC), which displayed a remarkable uniformity. No regional forms of the script could be discerned, although ethnic groups of varied cultural background throughout the vast expanse of the realm used it[8] (Naveh, 1972; Greenfield, 2001), the same script was used from Central Asia to Egypt, from the Caucasus to North Arabia (Greenfield, 1985, p. 709).

But after the fall of the Empire in the 3rd–2nd c. BC local varieties of the Old Aramaic script were developed in different cultural-geographic regions of the East, including Syria, North Mesopotamia,[9] Georgia and Armenia. Most forms of local Aramaic scripts began to crystallize in the first century BC.

Aramaic inscriptions of the South Caucasus found in Armenia[10] and mostly in Georgia clearly reflect this process (ibid., p. 702).

2. Studies on Old Aramaic Epigraphy in Georgia

The tradition of linguistic-paleographic studies of Old Aramaic epigraphy in Georgia is related to the name of the outstanding orientalist, Academician George (Giorgi) Tsereteli (1904–1973), who made a significant contribution to the decipherment and analysis of the Aramaic inscriptions discovered as a result of archaeological excavations at Armazi, near Mtskheta. To these inscriptions were devoted G. Tsereteli's two important works: *The Bilingual Inscription from Armazi* (1941, 1942) and *The Armazi Inscription of the Period of Mithridates the Iberian* (1962).

---

lary and terminology of the language, in which the preaching was performed and the canonical religious texts orally rendered. This, too, must have created the precondition for Georgian to have developed into a refined literary language by the time of the creation of Georgian Christian Script and its recording as a written language (Gamkrelidze, 1989, pp. 200–201).

8. On the Aramaic as an administrative language and *lingua franca* of the Persian Empire see also Folmer, 2011; 2020; Gzella, 2015b.

9. On the late Imperial Aramaic see Gzella, 2011; 2015a.

10. Old Aramaic inscriptions were discovered as well in Armenia, in different places, such as Zangezur, Teghut, Sevan, Sisian and Garni. They are inscribed on different objects: boundary stones, silver bowls, stones. Chronologically they could be attributed to 2nd c. BC-2nd c. AD. Their linguistic-paleographic studies were presented in the works of A. Perikhanian, 1964; 1965; 1966; 1971a,b.

As a result of paleographic studies of Mtskheta-Armazi inscriptions (11 lines bilingual (Greek-Aramaic) epitaphy, dated by 2nd c. AD (Fig. 2c)[11] and 14-line Aramaic monolingual inscription dated by 1st c. AD) (Fig. 2b), G. Tsereteli identified hitherto unknown type of Aramaic script as "independent branch of Semitic writing" and named it "Armazi Aramaic" according to the place of its finding.[12]

The bilingual Aramaic-Greek inscription was an epitaph of "Serapitis, the daughter of Zevakh the Younger, viceroy (*bṭḥš*) of King Parsman, wife of Iodmangan the victorious and winner of many victories, master of the court (ἐπίτροπος) of King Xsefarnug, [and] son of Agrippa, master of the court of King Parsman."

The second stele discovered near Mtskheta called Armazi Monolingual is known as the stele of victory of Sharagas, the viceroy of King Mithridates (1st c. AD).

The story in the text is told by Sharagas, who performed military operations, after the successful ending of which he reported to the great king Mithridates: "I gained this victory for you, my King".

These extensive Aramaic inscriptions were of great historical and cultural significance. Social titles, personal names, political events attested in them present the most valuable material for pre-Christian Georgian history.

G. Tsereteli distinguished a number of linguistic and paleographic peculiarities of the Armazian script conditioned by close cultural links with Ancient Aramaic-Iranian commonwealth on the one hand and with Hellenic cultural world on the other. Thus, while considering texts of bilingual and monolingual inscriptions, G. Tsereteli defined several similarities with contemporary Middle Iranian (Parthian, Middle Persian) and Semitic (Palmyrene) scripts. At the same time, Greek influence was also evident. For example, in Bilingual inscription using ʿ*ayin* for expressing *ē* in the proper name *Serapitis* (Aramaic סערפיט, Greek Σηραπειτίς/Σηραπιτίς[13]) is the early exemple of *mater lectionis*,[14] Iranian name Xšēfarnūγ (in the Aramaic text חסיפרנוג) was rendered by Greek form (*Ksyprnwg*), in Mithridates's inscription alongside with Aramaic *mlk*

---

11. The publication of the Armazi Bilingual attracted attention of many prominent Iranologists and Semitists. For the all reviews and notes on G. Tsereteli's work see G. Tsereteli, 1986, pp. 38–39, see also all scientific publications on this text: K. Tsereteli, 1992, pp. 115–118.

12. Both steles are kept at the Georgian National Museum, the Stone Fund, Bilingual (inventory number SSM 148) and Monolingual (SSM 149).

13. This transcription is given according to T. Kauchtschischwili, 2009, p. 390.

14. Another example of *mater lectionis* is found in Greek-Aramaic inscription on the silver spoon from Zghuderi. There are two graffiti, one is Greek: *XHΔ-Xηδ* and the other Aramaic: *kʿd-Ked*. It represents a complete or abbreviated name of the owner. The letter ʿ (ʿ*ayin*) was used as *mater lectionis* in Aramaic scribal tradition of Georgia (Chelidze, 1993, p. 21; Braund, 1994, 215, n. 64).

"the king" are attested Greek forms *bzys*, *bzls* (Βασιλεύς) and probably, Latin form *ḳysr* (Caesar) (G. Tsereteli, 1962, p. 375).

It is noteworthy, that local (Georgian) writing tradition was significantly reflected in the language and script of the Armazian inscriptions, namely, in similar outlines of several Armazian and Georgian letters, also transliteration and transcription of a number of Oriental terms and proper names, for example, Middle Iranian administrative name *bṭḫš* "a viceroy" presented in Armazian writings differently: *bṭḫš* (the bilingual inscription), *byṭy'ḫš* (the Bori inscription), which clearly reflects the impact of the Georgian orthography (cf. the Georgian *p'itiaxš-i*) (G. Tsereteli, 1948b, p. 56).

Armazian inscriptions attest distinguished forms of Middle Iranian proper names of Georgian nobles. These names reflect various dialect layers (south-western, north-western and north-eastern) and are mostly rendered according to their adequate pronunciation, without following the principles of Iranian historical orthography, for example, *Mhrdṭ*, *Myhrdṭ* (Monolingual), *'Sprwg* (Monolingual, cf. Ἀσπαρούνις on II century gem from Armazi, Georgian Asparug), *Šrgs* (Monolingual), *Bwz-Myhr* (Bori inscription, cf. *Burzen-Mihr* in 5th c. Georgian inscription in Palestine), *Ywdmngn* (Bilingual).

G. Tsereteli outlined several distinctive grammatical characteristics as well as irregularities of the bilingual text (the lack of a definite article, misuse of genders, the absence of the determinative state, the use of archaic prounoun *zy*), which is certainly a result of the local Aramaic writing tradition.[15]

In Armazian writing Eastern and Western elements were transformed on the ground of the native culture, creating most original linguistic and paleographical material (ibid., p. 56). G. Tsereteli also named Armazian script as "Georgian-Aramaic" or "Iberian-Aramaic" (G. Tsereteli, 1942, 51, n. 2).

Studies of the Aramaic inscriptions from Armazi were of special significance not only as a new source for the research of Eastern Aramaic writing and its ramifications, but also shed light on a number of cultural-historical problems of pre-Christian Iberia and its interrelations with Ancient Iran.

The tradition of using the Aramaic script in pre-Christian Georgia is closely connected to the problem of the origins of the Georgian alphabet. G. Tsereteli considered it in genetic relation with the Aramaic script (G. Tsereteli, 1948a; 1949).

---

15. About the linguistic peculiarities of the Armazi inscriptions—and particularly about the Bilingual—that could not be due to scribal mistakes and misspellings, see Kutscher and Naveh, 1970; Skjaervø, 1995, p. 291.

Apart from Serapitis and King Mithridates' steles, inscriptions on different objects found during the excavations at Mtskheta-Armazi were made in Armazi script.

In this respect, items found in Armazi necropoleis, notably golden plaques (2nd c. AD), silver plate of the pitiakhsh Bersuma, golden rings and bracelets (3rd−4th c. AD) are of a special interest due to their epigraphical value, as well as artistic quality. They reflect national artistic tradition together with contemporary Hellenistic and Oriental cultural style, including Iranian (Chubinashvili, 2007). (Fig. 3a−d).

G. Tsereteli has shown in his researches that the inscriptions found in Mtskheta-Armazi as well as some other epigraphical monuments of Georgia-Bori (2nd−3rd c. AD) (Fig. 3e and 4a), Urbnisi (2nd c. AD) (Fig. 4b), are done in the same "Armazian" script, which are distinguished with common paleographic features.

The inscriptions of Bori as stated by G. Tsereteli, showed a certain tendency to mannerism and stylization, "the lines are broken and in the break places sharp angles are formed," which could be due to material on which the inscription was made.

The oldest one is the monolingual inscription (1st c. AD) made in cursive, where letters have little (if any) distinction from each other, cf. identical are *k* and *n* letters; *r* and *b*; *ʿ* and *š*; *t* and *y* etc. The script of the bilingual text is more formal (Fig. 4c), in which all letters have clearly outlined forms. Letters of the monolingual inscription (Fig. 2b), are distinguished by more variations compared to letters of the bilingual text. A number of letters in the bilingual inscription, as of the later monument, are significantly different from the monolingual's (Fig. 4d).

Here it should also be noted that the writing of each mentioned monuments (1st−3rd c. AD) is characterized by certain specific paleographic features. We cannot come across absolutely identical writing of one and the same letters not only in different "Armazi" texts, but sometimes they cannot be attested within the same texts either. Certain variations of identical letters in the "Armazi" script monuments are quite acceptable, but they are very rare and fall within the general limits of the script.

In his later works G. Tsereteli assumed that the Armazian writing originated in "a variety" of the Aramaic script, which was spread in North-Eastern Mesopotamia (Assur, Hatra, Hassan-Kef, Sari) during the Hellenistic epoch.[16]

In 1961 in Garni (Armenia), an Aramaic stone-inscription was found. It was published in 1964, by Anahit Perikhanian (1964) and attributed to the 2nd c. AD. The writing of the inscription from Garni paleographically was the most similar from all the Aramaic scripts to that of the Armazian inscription (Fig. 5a). It became clear that the Armazian script

---

16. These issues were further tackled in the works of the German Semitist Oelsner, 1973, pp. 430−434; 1976.

(a)

(b)

(c)

(d)

(e)

FIGURE 3. (a) The Plate of Bersuma, photo reprinted from Chubinashvili (2007, Illustration 28). (b) Bracelets, photo reprinted from Chubinashvili (2007, Illustration 17). (c) Serapitis' necklace and pendant with ram head relief, photo reprinted from Chubinashvili (2007, Illustration 7). (d) A gem portrait, photo reprinted from Chubinashvili (2007, Illustration 4). (e) Bori plate, photo reprinted from Smirnov (1909, N 305, Table CXXI:12).

(a)

(b)

(c)

Transcription		Dedoplis Mindori Plates		Armazi Bilingual	Bori	Urbnisi
Hebrew	Latin	№№1,4	№№9,12,20,21,25			
א	'					
ב	b					
ג	g					
ד	d					
ה	h					
ו	w					
ז	z					
ח	ḥ					
ט	ṭ					
י	y					
כ	k					
ל	l					
מ	m					
נ	n					
ס	s					
ע	ʿ					
פ	p					
צ	ṣ					
ק	k					
ר	r					
ש	š					
ת	t					

(d)
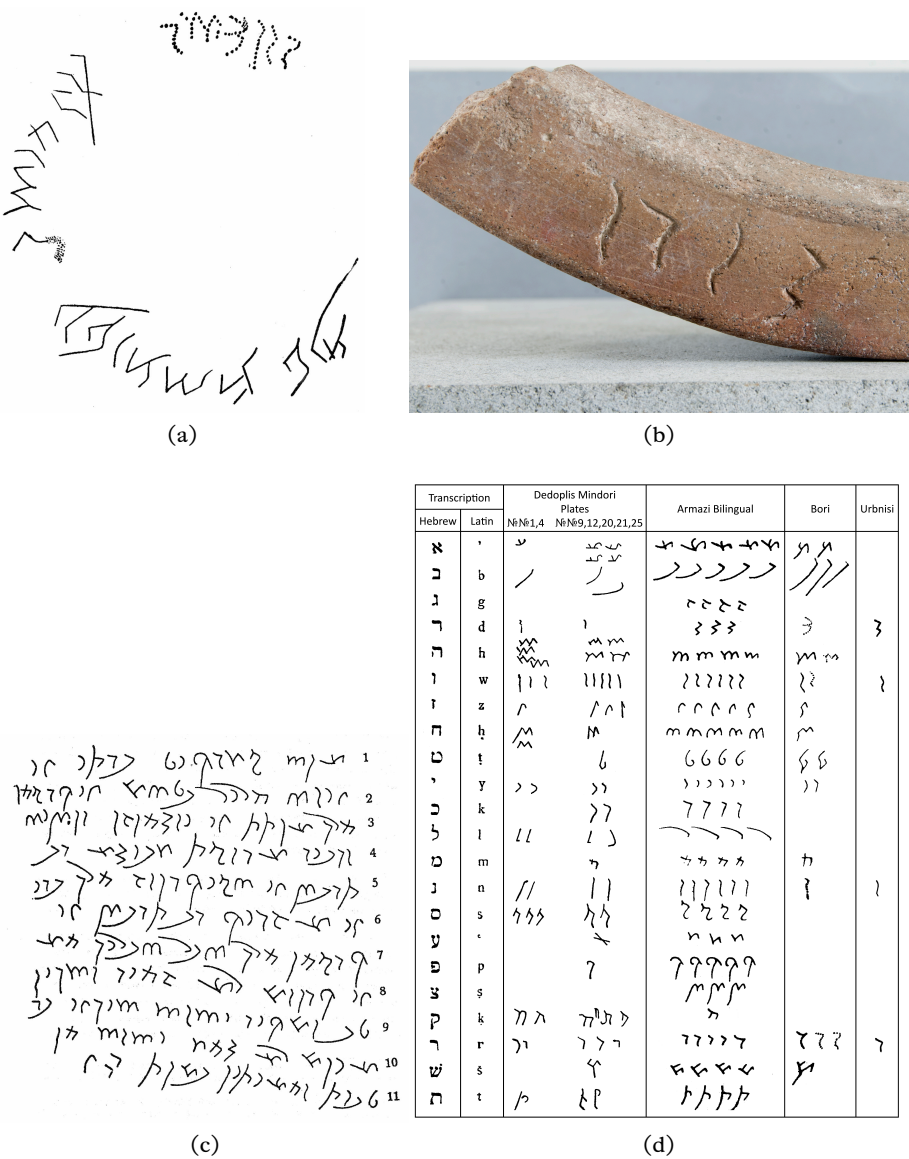
FIGURE 4. (a) A facsimile of the Bori plate inscriptions, reprinted from G. Tsereteli (1960, Table VI). (b) Aramaic inscription on the wine-cellar from Urbnisi, photo made at the Urbnisi Fund of the Georgian National Museum. (c) Armazi Bilingual's Aramaic inscription facsimile, reprinted from G. Tsereteli (1942, p. 15). (d) A table of Aramaic script types of Dedoplis Mindori plates (1st c. AD), Urbnisi inscription, reprinted from Gagoshidze and Tsotselia (1991, Addenda) and of Armazi Bilingual and Bori plate inscriptions, reprinted from G. Tsereteli (1948a, p. 100).

was characteristic not only of Georgian reality, but also of neighbouring Armenia. A. Perikhanyan suggested that the inscriptions on the Armazi steles, on the plate of Bori as well as the Garni inscription were written in the same script.

Aspects of comparative-historical development of the Armazian script were considered later by Joseph Naveh in the work "North Mesopotamian Aramaic script-type in the Late Parthian period" Naveh (1972), where paleographic analysis of Armazian letters and close to them (but not identical) letters of the Garni inscription was dealt in the common evolutionary typological scheme of Eastern Aramaic inscriptions of that period (such as Hatra, Dura-Europos, Hassan-Kef and others), several deviations of Armazian script from other writings were shown and the main tendencies of its evolution as an original type in North-Mesopotamian Aramaic writing branch were outlined (Fig. 5b).

The tradition of epigraphic Aramaic studies in Georgia was continued by another outstanding scholar, the late Professor Konstantin Tsereteli (1921–2004), who offered several works to newly-discovered Aramaic inscriptions (Uplistsikhe, 3rd–2nd c. BC) (Fig. 6a, 6b), Dedoplis Mindori (1st c. BC) (Fig. 6c).

In K. Tsereteli's works innovative theoretical assumptions were presented about the Georgian type of Aramaic script, by distinguishing three stages in its development: Pre-Armazian (Uplistsikhe inscriptions—this script was very close to the Official Aramaic and was considered as the predecessor of the Armazian (K. Tsereteli, 2001d), Early Armazian (Dedoplis Mindori inscriptions, which displayed more archaic features than later monuments, K. Tsereteli, 1993; 2001c), Armazian itself (Armazi steles, Urbnisi inscription, Bori silver plate inscription, etc.).

By considering rich factological material, K. Tsereteli defined common tendencies of the Armazian script type development in the South Caucasus: in the 3rd–2nd c. BC, a variety of the Old Aramaic script begins to be formed in this region and took its final shape in the 1st c. BC. In Georgia "Armazian" type of Aramaic writing (1st–3rd c. AD) was raised, typologically similar to the Aramaic script of Armenia but not wholly its identical. In both countries this type of writing was used before the adoption of Christianity (K. Tsereteli, 2001a).

3.   Future Prospects

Modern level of Old Aramaic Studies, new epigraphic findings and scientific publications essentially require a complex and systematic research of the Old Aramaic inscriptions of Georgia. The research will comprise two stages: (1) making a catalogue of edited inscriptions with their chronological distribution, photo material, texts, facsimiles, new

The Garni Inscription (Perikhanian's drawing)

(a)

A Comparative Script Chart:

Line 1. Hatra, No. 214
Line 2. Hatra, No. 35
Line 3. The Dura Europos
bilingual inscription
Line 4. Assur

Line 5. 'Abrat al-Ṣaghira
Line 6. Sari
Line 7. Hassan-Kef
Line 8. Garni
Line 9. Armazi

(b)

FIGURE 5. (a) The Garni inscription reprinted from J. Naveh (1972, p. 297). (b) A comparative chart of the North-Mesopotamian Aramaic script types, reprinted from Naveh (1972, p. 299).

(a)   (b)

(c)

FIGURE 6. (a), (b) Aramaic inscriptions on pieces of wine jars from Uplistsikhe, photos made at the Uplistsikhe Fund of the Georgian National Museum. (c) Aramaic inscription on a fragment of a wine pitcher from Dedoplis Mindori,photo taken at the Dedoplis Mindori Fund of the National Mesum of Georgia.

linguistic interpretations and comments together with a bibliographic index; (2) theoretical studies: a systematic linguistic-paleographic examination of published as well as unpublished material; their comparative analysis with Aramaic script of Armenia and other types of contemporary Eastern Aramaic writings; revealing paleographic peculiarities and evolutionary regularities of the South Caucasian Aramaic script.

The research will be essentially interdisciplinary, for the first time presenting the main tendencies of the Old Aramaic script's development in the light of Near Eastern—South Caucasian cultural-linguistic interference.

References

Academia Inscriptionum et Litterarum Humaniorum (1889). *Corpus Inscriptionum Semiticarum, Pars II, Tomus I*. Parisiis: E Reipublicæ Typographeo.

Andronikashvili, Mzia [ანდრონიკაშვილი, მზია] (1966). *ნარკვევები ირანულ-ქართული ენობრივი ურთიერთობიდან [Studies on Iranian-Georgian Linguistic Relations]*. თბილისი [Tbilisi]: თბილისის უნივერსიტეტის გამომცემლობა [Tbilisi Universiti Press].

Braund, David (1994). *Georgia in Antiquity: A History of Colchis and Transcaucasian Iberia, 550 BC-AD 562*. Oxford: Clarendon Press.

Chelidze, Maya (1993). "The Small Aramaic Inscriptions from the Village Zguderi." In: *Semitica. Serta philologica Constantino Tsereteli dicata*. Ed. by R. Contini, F.A. Pennacchietti, and M. Tosco. Turin: Silvio Zamorani Editore, pp. 15–22.

Chkheidze, Theo [ჩხეიძე, თეო] (1984). *ნარკვევები ირანული ონომასტიკიდან [Studies on Iranian Onomastics]*. თბილისი [Tbilisi]: მეცნიერება [Mecniereba].

Chubinashvili, Giorgi [ჩუბინაშვილი, გიორგი] (2007). *არმაზის პიტიახშთა ნეკრაპოლისში აღმოჩენილი ნივთების მხატვრული დახასიათების ცდა [Attempt of an Artistic Description of the Artifacts found in the Necropolis of Armazi]*. თბილისი [Tbilisi]: გამომცემლობა ნეკერი [Publishing House Nakeri].

Donner, Herbert and Wolfgang Röllig (1969). *Kanaanäische und aramäische Inschriften*. Vol. 3. Wiesbaden: Harrassowitz Verlag.

Folmer, Margaretha (2011). "Imperial Aramaic as an Administrative Language of the Achaemenid Period." In: *The Semitic Languages. International Handbook*. Ed. by Stefan Weninger et al. Berlin/Boston: Walter de Gruyter, pp. 587–598.

———— (2020). "Aramaic as Lingua Franca." In: *Companion to Ancient Near Eastern Languages*. Ed. by Rebecca Hasselbach-Andee. Blackwell Companions to the Ancient World. New York: John Wiley, pp. 373–399.

Furtwängler, Andreas et al., eds. (2008). *Iberia and Rome. The Excavations of the Palace at Dedoplis Gora and the Roman Influence of the Caucasian Kingdom of Iberia*. Vol. 13. Schriften des Zentrums für Archäologie und Kulturgeschichte des Schwarzmeerraumes. Langenweißbach: Beier & Beran.

Gagoshidze, Iulon [გაგოშიძე, იულონ] and Medea Tsotselia [მედეა წოწელია] (1991). "არამეულწარწერიანი ფირფიტები დედოფლის გორიდან [Plates with Aramaic Inscriptions from Dedoplis Gora]." In: *ამიერკავკასიის ისტორიის საკითხები [Issues of the History of Transcaucasia]*. თბილისი [Tbilisi]: მეცნიერება [Mecniereba], pp. 47–78.

Gamkrelidze, Thamaz [გამყრელიძე, თამაზ] (1989). წერის ანბაჭური სის-ტემა და ძველი ქართული დამწერლობა. ანბანური წერის ტიპოლო-გია და წარმომავლობა [*Alphabetic Writing and the the Old Georgian Script. A Typology and Provenance of Alphabetic Writing Systems*]. Ed. by Akaki Shanidze [აკაკი შანიძე]. თბილისი [Tbilisi]: თბილისის უნივერსიტეტის გამომცემლობა [Tbilisi Universiti Press].

————— (1994). *Alphabetic Writing and the Old Georgian Script. A Typology and Provenance of Alphabetic Writing Systems*. Caravan Books, Delmar-New York.

Gershevitch, Ilya (1979). "The Alloglottography of Old Persian." In: *Transactions of the Philological Society* 77.1, pp. 114–190.

Giorgadze, Grigol (2008). "The Armazian Script." In: *Iberia and Rome. The Excavations of the Palace at Dedoplis Gora and the Roman Influence of the Caucasian Kingdom of Iberia*. Vol. 13. Schriften des Zentrums für Archäologie und Kulturgeschichte des Schwarzmeerraumes. Langen-weißbach: Beier & Beran, pp. 253–257.

Greenfield, Jonas Carl (1985). "Aramaic in the Achaemenian Empire." In: *The Cambridge History of Iran, volume 2: The Median and Achaemenian Periods*. Ed. by Ilya Gershevitch, pp. 698–713.

————— (2001). "Aramaic." In: *Encyclopædia Iranica*. Vol. II. London-New York: Encyclopædia Iranica Foundation, pp. 251–252.

Gzella, Holger (2011). "Late Imperial Aramaic." In: *The Semitic Languages. International Handbook*. Ed. by Stefan Weninger et al. Berlin/Boston: Walter de Gruyter, pp. 598–609.

————— (2015a). "Aramaic in the Hellenistic and Early Roman Near East." In: *A Cultural History of Aramaic. From the Beginnings to the Advent of Islam. Handbook of Oriental Studies. Section I. The Near and Middle East*. Vol. 111. Leiden/Boston: Brill, pp. 212–279.

————— (2015b). "Official Aramaic and the Achaemenid Chancellery." In: *A Cultural History of Aramaic. From the Beginnings to the Advent of Islam. Handbook of Oriental Studies. Section I. The Near and Middle East*. Vol. 111. Leiden/Boston: Brill, pp. 157–208.

Kauchtschischwili, Tinatin [ყაუხჩიშვილი, თინათინ] (2009). საქარ-თველოს ბერძნული წარწერების კორპუსი [*Korpus der Griechischen Inschriften in Georgien*]. [Tbilisi]: ლოგოსი [Logos].

Kutscher, Edward Ychezkel [קוטשר, יחזקאל] and Joseph Naveh [יוסף נוה] (1970). "הכתובת הדו-לשונית (יוונית-ארמית) מארמאזי [The Bilingual Inscription from Armazi]." In: לשוננו [*Our Language*] 34, pp. 309–313.

Martirosyan, Hrach (to appear). *Iranian Personal Names in Armenian Collateral Tradition*. Iranisches Personennamenbuch. Vienna: Austrian Academy of Sciences.

Metreveli, Roin [მეტრეველი, როინ], ed. (2008). ქართლის ცხოვრება [*The Life of Kartli*]. [Tbilisi]: მერიდიანი არტანუჯი [Meridiani Artanuji].

Naveh, Joseph (1970). "The Development of the Aramaic Script." In: *Proceedings of the Israel Academy of Sciences and Humanities* 5.1.

Naveh, Joseph (1972). "The North Mesopotamian Aramaic Script-type in the Late Parthian Period." In: *Israel Oriental Studies* 2, pp. 293–304.

Oelsner, Joachim (1973). "Bemerkungen zur schriftgeschichtlichen Einordnung der Inschriften aus Armazi." In: *Wissenschaftliche Zeitschrift der Friedrich-Schiller-Universität Jena. Gesellschafts- und sprachwissenschaftliche Reihe* 22.3, pp. 429–438.

——— (1976). "Probleme der Entwicklung der aramäischen Schrift in Nordmesopotamien." In: *Philologia Orientalis* 4, pp. 215–223.

Perikhanian, Anahit [Периханян, Анаид Георгиевна] (1964). "Арамейская надпись из Гарни [Aramaic Inscription from Garni]." In: *Историко-филологический журнал, Ереван [Historical-Philological Journal, Erevan]* 3.26, pp. 123–138.

——— (1965). "Арамейская надпись из Зангезура (некоторые вопросы среднеиранской диалектологии) [Aramaic Inscription from Zangezur (Some Questions of Middle Iranian Dialectology)]." In: *Историко-филологический журнал, Ереван [Historical-Philological Journal, Erevan]* 4.31, pp. 107–128.

——— (1966). "Une inscription araméenne du roi Artašes trouvée à Zangézour (Siwnikʻ)." In: *Revue des études arméniennes* 3, pp. 17–29.

——— (1971a). "Inscription araméenne gravée sur une coupe d'argent trouvée à Sissian (Arménie)." In: *Revue des études arméniennes* 8, pp. 5–11.

——— (1971b). "Арамейская надпись на серебряной чаше из Сисиана [Aramaic Inscription on a Silver Cup from Sisian]." In: *Историко-филологический журнал, Ереван [Historical-Philological Journal, Erevan]* 3, pp. 78–82.

Qaukhchishvili, Simon [სიმონ, ყაუხჩიშვილი] (1955). *ქართლის ცხოვრება [The Life of Kartli]*. თბილისი [Tbilisi]: მეცნიერება [Mecniereba].

Rubio, Gonzalo (2006). "Writing in Another Tongue: Alloglottography in the Ancient Near East." In: *Margins of Writing. Origins of Cultures*. Ed. by Seth L. Sanders. Saline, MI: McNaughton & Gunn, pp. 33–66.

Shaked, Shaul (2006). "Notes on Some Jewish Aramaic Inscriptions from Georgia." In: *Jerusalem Studies in Arabic and Islam* 32, pp. 503–10.

Skjaervø, Prods Oktor (1995). "Aramaic in Iran." In: *Aram* 7, pp. 283–318.

Smirnov, Ya. I. [Смирнов, Я. И.] (1909). *Восточное серебро. Атласъ древней серебряной и золотой посуды восточнаго происхожденія, найденной преимущественно въ предѣлахъ Российской Имперіи [Oriental Silver. Atlas of ancient silver and gold dishes of oriental origin, found mainly within the Russian Empire]*. С.-Петербургъ [St. Petersburg]: Изданіе Имп. Археологической Коммиссіи [Imperial Archaeological Commission].

Tsereteli, George [Церетели, Георгий Васильевич] (1941). *Армазская билингва [Armazi Bilingual]*. Тбилиси [Tbilisi]: Издательство АН Грузинской ССР [Publishing House of the Academy of Sciences of Georgian SSR].

——————— (1942). არმაზის ბილინგვა [*A Bilingual Inscription from Armazi Near Mcheta in Georgia*]. Vol. 13. ნ.მარის სახელობის ენის, ისტორიის და მატერიალური კულტურის ინსტიტუტის მოამბე [The Bulletin of the Marr Institute of Language, History and Material Culture]. თბილისი [Tbilisi]: საქართველოს სსრ მეცნიერებათა აკადემიის გამომცემლობა [Publishing House of the Academy of Sciences of Georgian SSR].

——————— (1948a). "Армазское письмо и пробема происхождения грузинского алфавита. I [Armazian Script and the Problem of the Origin of the Georgian Alphabet". I]." In: *Эпиграфика Востока II [Epigraphy of the East II]*. Москва/Ленинград [Moscow/Leningrad]: Издательство академии наук СССР [Academy of Sciences of the USSR], pp. 90–101.

——————— (1948b). "Эпиграфические находки в Мцхета—древней столице Грузии [Epigraphic Discoveries in Mtskheta—the Ancient Capital of Georgia]." In: *Вестник Древней Истории [Bulletin of Ancient History]* 2.24, pp. 49–57.

——————— (1949). "Армазское письмо и пробема происхождения грузинского алфавита. II [Armazian Script and the Problem of the Origin of the Georgian Alphabet". II]." In: *Эпиграфика Востока III [Epigraphy of the East III]*. Москва/Ленинград [Moscow/Leningrad]: Издательство академии наук СССР [Academy of Sciences of the USSR], pp. 59–71.

——————— (1960). "Древнейшие грузинские надписи из Палестины [The Most Ancient Georgian Inscriptions from Palestine]." In: Тбилиси [Tbilisi]: Издательство академии наук Груз. ССР [Academy of Sciences of the Georgian SSR].

——————— (1962). "Армазская надпись эпохи Митридата Иверийского [The Armazi Inscription of the Period of Mithridate the Iberian]." In: *Труды 25-го Международного конгресса востоковедов [Proceedings of the XXV International Congress of Orientalists]*. Москва [Moscow], pp. 374–378.

——————— (1974). "The Achaemenid State and World Civilization." In: *Acta Iranica* 1, pp. 102–107.

——————— (1986). ბიობიბლიოგრაფია / Биобиблиография [*Biobibliography*]. თბილისი / Тбилиси [Tbilisi]: მეცნიერება / Мецниереба [Mecniereba]. URL: http://science.org.ge/old/members/BioBibliografia/Wereteli%5C%20Giorgi.pdf.

Tsereteli, Konstantin [კონსტანტინე, წერეთელი] (1976). "სირიელისა და ასურელის აღმნიშვნელი ტერმინები ქართულში [Ethnic Terms denoting "Syrian" and "Assyrian" in Old Georgian]." In: *ივანე ჯავახიშვილის დაბადების 100 წლისადმი მიძღვნილი საიუბილეო კრებული [Volume of articles dedicated to the 100th anniversary of Academic Ivane Javakhishvili]*. თბილისი [Tbilisi]: მეცნიერება [Mecniereba], pp. 177–188.

Tsereteli, Konstantin [კონსტანტინე, წერეთელი] (1992). *შენიშვნები არ-მაზის ბილინგვის ტექსტზე [Notes on the Text of Armazi Bilingual]*. თბი-ლისი [Tbilisi]: მეცნიერება [Mecniereba].
—————— (1993). "The Oldest Armazian Inscription in Georgia." In: *Die Welt des Orients* 24, pp. 85–88.
—————— (1994). "Aramaic Language in Georgia." In: *Proceedings of the Eleventh World Congress of Jewish Studies, Division D*. Vol. I. Jerusalem: Magnes Press, pp. 9–16.
—————— (1996). "An Aramaic Amulet from Mtskheta." In: *Ancient Civilizations from Scythia to Siberia* 3.2–3, pp. 218–240.
—————— (1998a). "Epitaph des Jehuda Gurk." In: *Georgica* 21, pp. 74–78. The author's name is spelled "Zereteli".
—————— (1998b). "Les inscriptions araméennes de Géorgie." In: *Semitica* 48, pp. 75–78.
—————— (2001a). "Armazian Script." In: *სემიტოლოგიური და ქართვე-ლოლოგიური შტუდიები [Semitological and Kartvelological Studies]*. თბი-ლისი [Tbilisi]: ლოგოსი [Logos], pp. 419–427.
—————— (2001b). "Die alte aramäische Inschrift aus Georgien." In: *სე-მიტოლოგიური და ქართველოლოგიური შტუდიები [Semitological and Kartvelological Studies]*. first published in *Brücken, Festgabe für Gert Hummel zum 60. Geburtstag*, Tbilissi, 1993. თბილისი [Tbilisi]: ლოგოსი [Logos], pp. 385–392. The author's name is spelled "Zereteli".
—————— (2001c). "ორი მცირე არამეული წარწერა დედოფლის გორიდან [Two Short Aramaic Inscriptions from Dedoplis Gora]." In: *სე-მიტოლოგიური და ქართველოლოგიური შტუდიები [Semitological and Kartvelological Studies]*. თბილისი [Tbilisi]: ლოგოსი [Logos], pp. 381–384.
—————— (2001d). "უფლისციხის არამეული წარწერები [Aramaic Inscriptions from Uplistsikhe]." In: *სემიტოლოგიური და ქართველოლოგი-ური შტუდიები [Semitological and Kartvelological Studies]*. თბილისი [Tbilisi]: ლოგოსი [Logos], pp. 343–363.

On the Typology of Writing Systems

Liudmila L. Fedorova

Abstract. The paper aims to propose a scheme for classification of writing systems, based on four binary characteristics of spelling: linear vs. non-linear spelling, integral vs. segmental one, complete vs. reduced, simple vs. differentiated spelling. The main attention is given to the non-linear emblematic writing, namely to Aztec script, which shows the examples of linguistic emblems—the first readable writing signs for place names and proper names. Further development of writing explores the techniques of segmentation and differentiation that contribute to the refinement of spelling, yet they go along with a trend to reduced and integrated forms, so we have today the coexistence of emblems-emoticons, Chinese characters and highly differentiated alphabets.

1.   Introduction. The Problem of Typology of Writing Systems

The aim of the present paper is to demonstrate how the existing typology of writing systems can be further refined using additional criteria for classification based on the main capabilities of a writer and a reader to compose and decompose, to integrate and to differentiate.

When speaking about historical scripts, it is necessary to distinguish between, on the one hand, the first attempts of using graphic images and signs and, on the other hand, writing practices that have been developed, based on systems of signs. This division was firstly established by E. Taylor, who distinguished two stages, corresponding to pictography and phonography, the former considered as 'proto-writing', and the latter as 'true writing'.

The proto-writing stage is nevertheless not reduced to pictography alone, and the evolution of writing does not always correspond to the widespread cliché 'from picture to letter'.

Liudmila L. Fedorova   ⓘ 0000-0002-2284-6643
Russian State University for the Humanities,
119421 Russia, Moscow, Obrucheva 28-8-105
E-mail: lfvoux@yandex.ru

First of all, it should be noted that graphic signs can be used for different purposes and therefore for different functions: magic, social, cognitive, mnemonic, decorative, etc., so that the communicative function (transmission of messages) has been only one among others and obviously not the first one. The use of graphic (or painted) marks in many cases does not differ substantially from the use of object signs such as amulets, counting tokens, status attributes, etc. Totems and amulets as signs of upper patronage could be objects or also graphic or painted marks (e.g., a handprint on a cave wall), and similarly for signs of social status, of self-identification (e.g., tattoos), of property (tamgas), of association or 'affiliation', of contract, of authenticity, of war or peace; they could form their own symbolic systems of objects and marks, made of distinct graphic images and 'empty' figures or lines without visible reference. So the variety of possible finalities of graphic signs should be taken into account in the investigation of the beginnings of writing per se. The main distinction of writing signs from other marks is not their form, but their function and the way they are used: for transmission of information, for communication or just for demonstrative purposes.

Therefore writing systems can be regarded as a case of a more wide class of semiotic systems, with their own tasks and ways of functioning.

When writing systems are conceived as linguistic systems, they are defined according to their phonetic values and to their capability of transmitting speech. While signs of most semiotic systems can be only interpreted, signs of linguistic writing systems can be *read*, they refer to language units.

The first classification of phonographic writing systems was proposed in the 19th century in the works of I. Taylor ('The Alphabet' 1883, cf. Daniels, 1996)), it distinguishes logographic, syllabic and alphabetic systems. This division, though rather speculative, remains a convenient scheme and a starting point for more detailed classifications. Further contributions to the study and systematization of writing systems have been made by J. Friedrich, D. Diringer, C. Loukotka, I.J. Gelb, V.A. Istrin, and others. At present, there are various classifications of writing systems that examine in detail the relationship between writing units and language units. These are works by J. Sampson, J. DeFrancis, W. Bright, R. Sproat, P. Daniels, F. Coulmas, H. Rogers, M. Neef, and others.

To return to the original classification, scholars admit that most writing systems have a mixed nature; first of all this concerns 'logographic' systems, for usually they include both ideographic and phonographic (mostly syllabic) signs. Ideography deals with the level of notions, which may or may not correspond to definite single words: an ideogram may correspond to a space of synonyms or related nouns, or to a class of words with the same root morpheme. So the first class of writing sys-

tems may be called logo-syllabic or morpho-syllabic. Yet the reference of 'logo-' or 'morpho-' items is controversial.

The syllabic class also seems to be heterogeneous. Gelb distinguished 'Aegean' systems as a specific class using signs for open short syllables (Gelb, 1963). These are qualified as being based on moras, so this type was later called *moraic*. Another subclass includes brahmi, devanagari and other derived systems that use specific operational techniques of vocalization; these were qualified as alphasyllabary (Bright, 2000) or abugida (Daniels, 2009a,b). In the Egyptian hieroglyphic script the subsystems of consonant 'alphabet' and 2-/3-consonant characters were regarded by Gelb as syllabic, due to the pronunciation practice. This also allowed qualifying some other West Semitic scripts as not consonant alphabets, but as a special type that was later named abjad (Daniels, 2009a,b).

The class of alphabets turned out to be heterogeneous as well. The Korean alphabet with its codification of articulation in parts of characters was qualified by P. Daniels as a 'featural' alphabet.

So the original classification evolved into something more complicated. Scholars proposed their own classification schemes with regard to different criteria of categorization.

2.   Classification of Writing Systems by H. Rogers

An important generalization was made by H. Rogers, who took into account three main dimensions of writing systems: (1) type of phonography, (2) amount of morphography, and (3) orthographic depth. It is represented in Scheme 1.

1. The *type of phonography* is given in the horizontal dimension of the scheme: abjad, alphabetic, abugida, moraic, syllabic.
2. The *amount of morphography* is given in the vertical dimension); 'it is higher when there are symbols that represent the morphemes' (e.g., <7>, <$>), or 'when spelling distinguishes morphemes (<by>, <bye>, <buy>)'.
3. *Orthographic depth*, which is greater when homophonous allomorphs are spelled similarly (*child—children*, *sign—signal*); it is denoted by the choice between uppercase and lowercase characters in the scheme.

There are five types of writing systems in this classification. The writing systems of different languages in every type can also be characterized by two gradual properties of spelling.

As can be seen in Scheme 1, some languages are located between classes, such as Sumerian, located between moraic and syllabic writing, and Pahawh Hmong, located between alphabetic and abugida writing. Rogers assumes that there is no clear division by class, but rather

	Abjad	Alphabetic		Abugida	Moraic	Syllabic
Amount of morphography	W. Semitic	Finnish Greek Belorusian	Pahawh Hmong	*Devanāgāri* BURMESE TIBETAN	Linear H CHEROKEE	Modern Yi
		KOREAN RUSSIAN SCOTS GAELIC				
	Perso-Aramaic	ENGLISH				
						Chinese
	Egyptian				Mayan Japanese Sumerian	
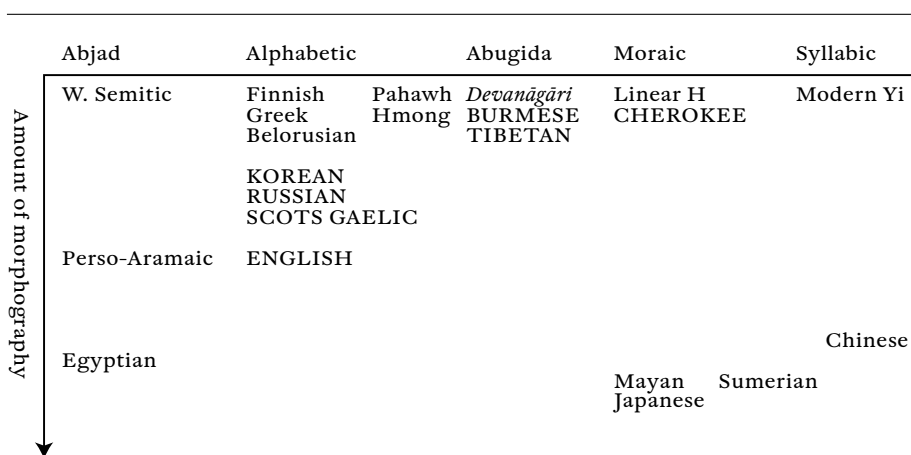
SCHEME 1. Rogers' generalized classification (Rogers, 2005, p. 274) (uppercase letters denote deep systems; lowercase letters, shallow ones)

a continuous space, in which some scripts cannot be clearly assigned to a specific class. This may indicate that they have properties of different classes, or that the criteria for determining them have not been developed. Rogers argues that this abstract space is a proper representation for the scheme, because many scripts are of mixed nature.

Still there arise some questions about this classification.

Why are moraic systems opposed to abjad and abugida? The nature of basic syllables may be also moraic in these systems for they have additional means to designate long vowels in a syllable—by two moras, or by 'weak consonants' as *matres lectionis* in abjad, or by distinguishing the diacritics for short and long vowels in abugida.

The definition of the amount of morphography presupposes two different cases. Is it always a matter of morphography, when semantic units are given as single signs (<7>, <%>)? Is it more convenient to speak about the amount of ideography, which can be defined by the number of ideograms? Rogers does not use the term of 'ideogram' because of its ambiguity, he rather refers to 'abstract pictograms'. Indeed, ideograms are usually opposed to pictograms, the former referring to the more elaborated type of writing than pure pictography, but these terms—'pictogram' and 'ideogram'—are not really opposed. While *picto*-gram refers to *signans*, the pictorial form of a sign, *ideo*-gram presupposes its content, 'idea', *signatum*. So an ideogram can very well take the form of a pictogram, so that the opposition between them vanishes. V.A. Istrin speaks about *fraseography* in both cases, distinguishing pictograms and abstract symbols (Istrin, 1965). Nevertheless we do not refrain from using the term of 'ideogram' for a written sign; we use it for a linear sign or for a pictogram, when it refers to an abstract notion on the base of

semantic shift. In turn, the term of 'pictogram' is more appropriate for the iconic image in its literal sense (a pictogram $<\maltese>$ can literally denote the sun, but as an ideogram it can have meanings such as 'light', 'day' based on a metonymic shift, or the meaning of 'majesty' through a metaphoric shift).

3.   An Additional Categorization

Let us return to the first classification of writing systems in its widely accepted form, and develop it by adding further divisions. The logic of dividing classes presupposes binary branching on the hierarchical levels. As a result we obtain eleven subclasses, labeled by traditional or mostly representative labels; some subclasses are attested only in a single writing system example and therefore are labeled by its name.

Three traditional classes can be distinguished with respect to the type of phonography: morphosyllabic (or logosyllabic), syllabic, and alphabetic writing, each one having its own subtypes (subclasses):

(A) morphosyllabic/logosyllabic writing is mixed, with two types of graphemes: morphemes/words (semantic units) or phonetic segments of syllabic type:
  (1) nonlinear systems (mixed emblematic type);
  (2) linear systems (mixed linear type).
(B) Syllabic writing:
  (1) primal syllabic (integral) spelling with graphemes, corresponding to syllables or phonetic segments of syllable types (CVC, CCVC, and CV, CVV, CVCV...; there may be more than one syllable in a grapheme);
    (a) its complete form is represented in many ancient scripts; in the modern *lolo* writing system, more than 800 graphemes are used to represent all possible syllables (Bradley, 2009);
    (b) its reduced (non-vocalized) form is given in Egyptian hieroglyphic (polyconsonantal) writing;
  (2) moraic kana-type writing, with graphemes denoting indivisible phonetic syllables or segments (CV, V, -C); examples are Aegean scripts in the ancient world and kana systems in modern Japanese;
  (3) moraic abugida writing with a standard subsystem of vowel modifications (C^V, V); examples are Indian scripts and their derivatives, as well as the Ethiopic script;
  (4) moraic reduced writing, abjad: graphically non-vocalized type, but based on vocalized units in pronunciation, presupposing an indefinite vowel in syllables (C^x); examples are West Semitic scripts;

(C) Alphabetic writing, where the main character/letter, corresponds to a sound/phoneme:

  (1) non-vocalized (reduced) writing, in which only consonants are independent graphemes (consonantic alphabet); modern Arabic;

  (2) linear writing, with vowels and consonants sequentially written as equal independent graphemes (linear alphabet, also with possible diacritic differentiation); Greek, Cyrillic, Latin, Armenian, and others;

  (3) nonlinear writing, with vowels and consonants written in inverted order (Pahawh Hmong is the only known example);

  (4) featural nonlinear writing with graphemes constructed through elements that differentiate articulation features of phonemes (featural Korean, cf. Daniels, 1996; Lee, 2009).

Morphosyllabic nonlinear systems are the most elementary examples of information recording by means of composition of signs, as linguistic emblems (using rebus spelling). In the Aztec script such records convey only some nominations—usually place names or personal names as readable emblems.

Morphosyllabic linear systems already convey a sequence of reading signs, for words and syllables, although they may allow some violations of the linear order (for example, the ornamental arrangement of signs in Mayan spelling, or the 'honorifical' order in Egyptian spelling, or graphic blocks in Chinese). This is a general phenomenon, observed in elaborated ancient scripts. The historical morphosyllabic systems usually have a rather representative and stable class of ideograms for semantic units and a more compact class of syllable signs.

The syllabic component of morphosyllabic systems can be further analyzed with respect to the organization of pronunciation units (type of phonography). Many syllabic systems have evolved historically out of morphosyllabic in order to get adapted to different languages. Whole-syllable (integral) spelling is opposed to moraic spelling: the former uses indivisible units while the latter uses decomposed, segmental ones. Moraic systems have their own subclasses: kana, abugida and abjad. We consider abjad as a moraic system for its characters presume vocalized consonants as minimal pronunciation units, naturally used in spelling and reading.

Alphabetic systems can be complete (vocalized) or reduced (consonant). The latter are systems derived from abjad using diacritics for vocalization. Their characters are not syllables anymore because vowels have their own representations.

Linear alphabets can be based on simple characters or include characters differentiated by diacritics.

"Non-linear alphabets" are those in which the order of graphemes within a syllable is violated (while the order of syllables is linear; so

they use both linear and non-linear order): this is the case of the Pahawh script where graphemes in a syllable are displayed in reverse order, and in Korean syllable blocks. These types have unique representations.

This classification is displayed in Scheme 2.
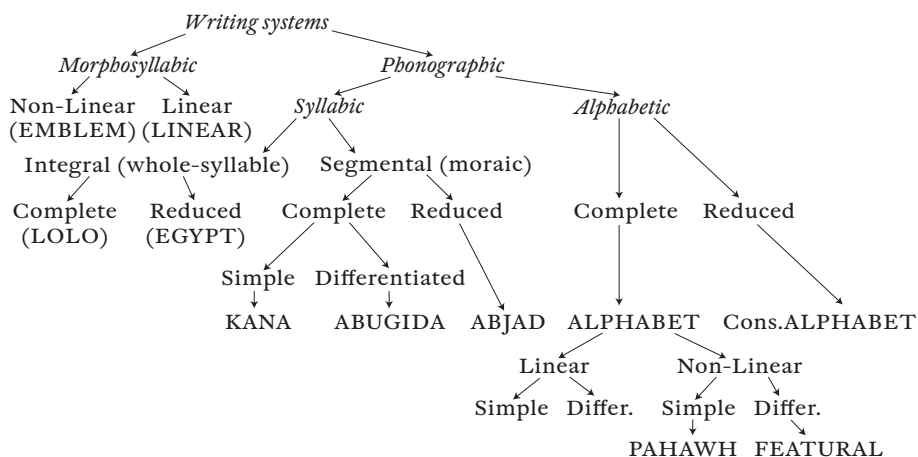
Scheme 2. The revised classification

As a result we have three main classes (MORPHOSYLLABIC, SYLLABIC, ALPHABETIC) and their subclasses (EMBLEMatic, LINEAR morphosyllabic, LOLO, EGYPTIAN, KANA, ABUGIDA, ABJAD, ALPHABET, CONSONANT ALPHABET, and also two unique "non-linear Alphabets"—PAHAWH and FEATURAL KOREAN. It should be taken into account that morphosyllabic types can be qualified as mixed systems with syllabic components (e.g., Linear B can be qualified as a morphosyllabic writing system with a kana-type syllabic component).

It can be seen that the categorization is based on four binary characteristics of spelling:

1. *linear/nonlinear spelling*: ex. g.: <1 - 2 - 3 - 4> vs. <2^2 - x_3>;
2. *integral (whole-syllable)/segmental (moraic)* spelling: [CCVC], [CVCVC] vs. [CV]-[CV]-[CV];
3. *complete (vocalized)/reduced (consonantic)* spelling: [CV] vs. [C^x];
4. *simple/differentiated* spelling: [CV] vs. [C^v].

These binary oppositions can operate at different levels of analysis that allow more detailed classification.

Let us examine them more carefully.

3.1. Linear vs. Nonlinear Spelling

Linearity is the first significant dimension of classification. Linear arrangement is an important step in the formation of phonetic writing. It follows the deployment of speech in time using one graphic dimension—on a line, be it horizontal or vertical. It is opposed to a non-linear, emblematic layout of readable graphic units which appears at the first stage of logo-/morpho-syllabic writing. Emblematic writing is in turn opposed to pictography and ideography where glyphs are non-readable signs and images, and just interpreted symbols. Linguistic (readable) emblems first appear in a pictographic frame representation for rendering names and numbers that correspond to words. While we have a single sign, the fact whether it represents a notion or a concrete word is ambiguous, be it an ideogram or a logogram. Only names can be phonetically reconstructed, and only in the case when they are represented as composition of signs with rebus spelling.

Yet readable emblems usually have reduced representations, for they allow only partial reading, using rebus spelling and omitting some elements. Their use can be observed in the Aztec codices that combine pictographic and phonographic techniques.

3.1.1. *Aztec Emblems in the Space of Pictorial Text*

The term 'emblem' was firstly introduced in the investigation of writing systems by H. Berlin (1958, pp. 111–119), yet not in the linguistic sense. It was used for signs designating Maya place names, which Berlin presupposed to be not readable, but only requiring interpretation. Place names got their readings in the decipherment of Maya script by Ju. V. Knorozov. In Aztec manuscripts, place-name emblems are also readable signs, though they have pictorial form and are used in a pictographic context, where events are represented by iconic images. The term of 'emblematic writing' has been introduced in Fedorova (2009), along with the notion of *linguistic emblem*.

The use of linguistic emblems can be illustrated by examples from Codex Mendoza, an Aztec manuscript, written in 1547, edited and commented by F. Berdan (Berdan 1997). My analysis is based on the Berdan's comments, on the *Nahuatl Grammar* by T. Sullivan (1983) and *Nahuatl Dictionary* by Rémi Siméon (1857) edited online[1] by Alex Wimmer. Its first part is a chronicle.

The beginning of Codex Mendoza (Fig. 1a) is consecrated to the foundation of Tenochtitlan. It uses the stable arrangement of pictorial glyphs: the central symbol indicates the main subject of the narrative

---

1. Bodleian Library, Oxford UK, `https://digital.bodleian.ox.ac.uk/objects/` `2fea788e-2aa2-4f08-b6d9-648c00486220`
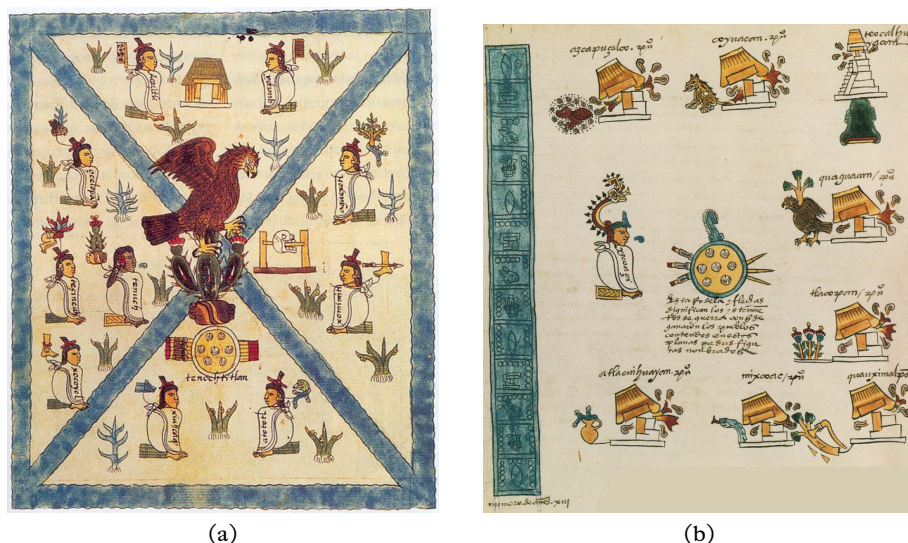
FIGURE 1.   (a) Foundation of Tenochtitlan (Codex Mendoza, 2ʳ, fragment). (b) The conquests of Lord Itzcoatl (Codex Mendoza, 5ᵛ)

event, the surrounding glyphs designate its participants, and the marginal frame serves for calendar emblems. The standard frame representation can convey information that may correspond to text; we may use the term '*textogram*' for it, following I. M. Dyakonov (1976, p. 570). Only some emblems can be read.

The central image is an emblem of Tenochtitlan, it is composed of three glyphs: a stone (*te-tl*), a cactus (*noch-tli*) on it (absolutive suffixes *-tl*, *-tli*, *-li* do not participate in compounding), and an eagle in the middle of the cactus to convey the sense 'among' (*-titlan*). The whole composition is based on another emblem: a shield with arrows, as a symbol of war; it indicates the conquest of the territory. The symbolic emblem of war is not readable (it is understood without reading), though there is a stable binomial expression in Nahuatl, *mitl chimalli* 'arrows, shield', which could correspond to it. So two root components of *Te-noch-titlan* can be read, and the locative suffix can be reconstructed. We do not know whether the etymology of the name should be understood as 'a place of cactus among stones'; more probably it refers to the name of a founder of Tenochtitlan, Tenoch, and should be read 'among the people of Tenoch'. The scribe may also be using a visual image that has a readable parallel as a rebus, so we can consider it as a linguistic emblem. The role of the eagle is not only semantic, but primarily symbolic, for according to a prophecy by wise men, the city was to be founded at the place where

an eagle would sit. So the scribe used iconic images with phonetic and symbolic values.

Other readable emblems are the Lords' names that are attached to the pictorial glyphs—standard images of Lords.

Another textogram example can be seen in Fig. 1b. It is dedicated to the conquests of Lord Itzcoatl. Lord Izcoatl ('snake with arrows' *coa-tl* 'snake', *iz-tli* 'arrow'), a name-emblem attached to his head, 'speaks' (a blue scroll as a sign of speech) about his war conquests (emblem of war—the shield and arrows), which are given in the emblems of a 'conquered city'—a burning and falling temple. Each city-emblem has an attachment that renders its name: the name's emblem. The whole can be interpreted: *Izcoatl speaks: I have conquered these cities...* It should be noted that the word for Lord *tlahtoani* literally means 'speaking' in Nahuatl, so the scroll may serve as a status indication.

3.1.2.   Examples of Linguistic Emblems in Nahuatl

The arrangement of readable name-emblems is non-linear, it is a composition of images that can represent an imaginary scene. Here are some examples.

The emblem of CUAUH-NAUAC resembling to a "speaking tree" (Fig. 2a) represents *cuahu-itl* 'tree' + *nahua-tl* 'speech', homophone of locative suffix *nahuac* 'near'; the resulting meaning is 'near trees'.

The emblem of AHUACA-TLAN "tree with teeth" (Fig. 2b) stands for *ahuaca-tl* 'avocado' + *tlan-tli* 'teeth', homophone of locative suffix *tlan* 'where there is a lot of...', 'among...' to express the sense of 'the place, where there is a lot of avocado trees' (TREE and AVOCADO use similar glyphs, but a reader could recognize compound names). Both cases are examples of rebus substitution.

Figures 2c and d show another way of phonetic representation, using phonetic complementation, a hint given by rebus reduplication. There we have two versions of the same place-name emblem of CUA-HUAH-CAN: *cuāuh-tli* 'eagle' reduplicated by *cuahu-itl* 'tree' (Fig. 2c), or: *cua-itl* 'head' of *cuāuh-tli* 'eagle' (in one graphic image) reduplicated by *cuahu-itl* 'tree' (Fig. 2d); the next two components have no visual expression: *huah* (possessive suffix) + *can* (locative suffix); the whole designates 'the place of owners of eagles', or 'the place of eagles'. Fig. 2c shows the name-emblem attached to the symbol of burning temple that means 'conquered city', Fig. 2d represents the same name bound to the glyph HILL (*tepe-tl*) for 'city, settlement' (*altepe-tl*). Symbols of BURNING TEMPLE and HILL are just pictorial images, they serve as a base for linguistic emblems.

Place-name emblems usually are attached to emblems of cities or burning temples, but they can also be used independently, designating tribes or settlements in the lists of tributes.
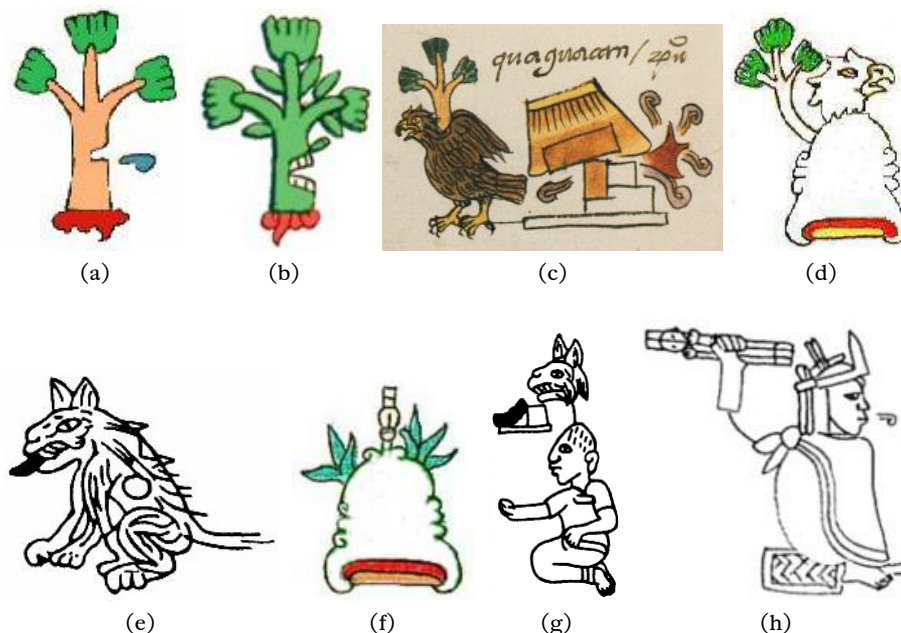
FIGURE 2.   (a) CUAUHNAUAC; (b) AHUACATLAN; (c) (conquered city of) CUAHUAHCAN (quaguacan); (d) (city of) CUAHUAHCAN; (e) COYUCAN; (f)ACATEPEC; (g) COYUCAC; (h) ACAMAPICHTLI

Fig. 2e, COYU-CAN, 'the place of (lean) coyotes' (or COYU-HUAH-CAN 'the place of owners of coyotes'), shows another example of phonetic complementation: *coyo-tl* 'coyote' with a round hole *coyoc-tli* 'hole' or *coyoc-tic* 'hole ridden'; the phonetic hint confirms the meaning 'coyote' (not 'dog'). The locative suffix *can* has no visual expression.

Fig. 2f, ACA-TEPE-C, with the morphemic structure: *aca-tl* 'reed' + *tepe-tl* 'hill' + *c* (locative suffix) 'on the hill of reeds', seems to be a direct iconic image of the name. Yet it uses a special semantic hint to confirm its reading. Its visual representation includes three components: hill (*tepetl*), reed (*acatl*) and dart (*acatl*). The glyph of hill is readable and serves also as a graphic base for other symbols. The new phonetic device is the semantic reduplication for *acatl*: it is given in two images: grass and a dart, corresponding to the meaning of *acatl*, and knowing that a dart is made using a reed's stem. It serves to recognize the image of reed that otherwise could be understood as plain grass or an arbitrary plant. The locative suffix has no visual representation, though its meaning can be implicitly assumed from the arrangement of two small glyphs on the top of the big one (HILL).

Figs. 2g and h represent the names of the tribe COYU-CAC and of Lord ACA-MAPICH-TLI. These names are attached to images of persons. The name COYU-CAC is divided in parts in order to provide a rebus representation: *coyo-tl* 'coyote'+ *cac-tli* 'sandal'; this is a decomposed rebus spelling (under the hypothesis of a rebus substitution for both parts of the word). The Lord's name probably represents its content in graphic images: *aca-tl* 'dart' and *mapich-tli* 'hand, fist' that means 'a fist holding darts'. It seems like a rather iconic representation, yet for the native readers these images refer to concrete words for 'fist' (not arm) and 'darts'. In fact, in this name, two principles of writing coexist: ideographic, presupposing reference to a notion, and phonographic, referring to a word. We may be confident in the phonographic nature of this sign, since it is confirmed by rebus spelling; yet we may suggest that images in name-emblems were recognized by native users in their exact phonetic form, as words, because the combination of glyphs increases the chance of guessing their fixed phonetic forms corresponding to a name.

The complexity and ingenuity of Aztec script consist in the decomposition of whole names and in the use of the same glyphs for pictographic and phonographic functions.

3.1.3. *Graphic Arrangement of an Emblem*

The previous examples show that an Aztec linguistic emblem usually consists of two (or three) meaningful graphic components, which are sufficient for the reconstruction of the whole name.

The arrangement of readable components relies on a decision taken by the scribe. Locative suffixes can be transferred by mutual disposition of components, as in TENOCHTITLAN and ACATEPEC. The whole composition can be done in different ways: by syncretism or reduplication, incompletely or by reduplication.

       (a)   (b)   (c)
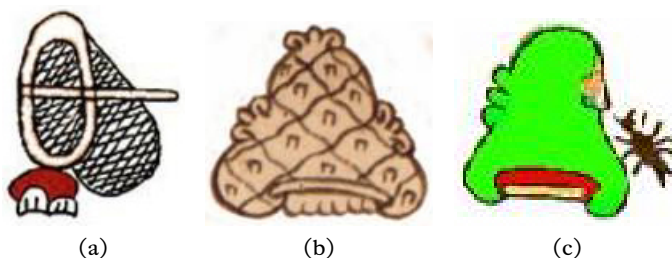
FIGURE 3. (a) MATLATLAN (Codex Mendoza). (b) MATLATLAN (Historia Tolteca Chichimeca). (c) YACAPICHTLAN (Codex Mendoza)

Thus, the place-name 'Matla-tlan' ('net' + 'where there is a lot of...' or 'in') is given differently in the two codices: in juxtaposition of a net *matlatl* and teeth *tlantli* (Fig. 3a) or as a hill *tepetl* (non-readable base) in a net.

The place-name 'Yacapich-tlan' (*yacapitz(-ahuac/-actic)* 'pointed' + *tlan* 'where there is a lot of...', together: 'the place of many pointed things') is represented as a hill with a nose, *yaca-tl* (for 'pointed') and an insect, bug *petz(o-tli)* which can bite it; the scribe's witty invention.

The logical incompatibility of images or, in the contrary, their fantastic combination into an entire image, and the incompleteness of spelling are characteristic of the graphic display of a linguistic emblem. They presuppose 'the intuition of meaning' or 'feeling of meaning', which J. Elkins qualifies as nesessary for understanding any sort of emblem (Elkins, 2003).

It follows that the main writing techniques in Aztec emblems consist in rebus substitution, rebus phonetic complementation (phonetic reduplication), decomposed rebus spelling, and semantic reduplication (semantic-phonetic analogy), when the scribe provides two parallel images corresponding to different meanings of a polysemic word (not homonyms). The scribe may combine direct iconicity and language game, phonetic analogy and semantic hint in a composition that corresponds to a compound word.

Thus we can define a linguistic emblem as a readable complex sign with a linguistic referent. Its main properties are the function of nomination (usually proper names and place names), non-linear arrangement of components, their limited number (usually 2-3), a new meaning of the whole that is not just a sum of meanings of its components, and therefore the possible incompleteness of spelling. For, as W. von Humboldt noted, synthesis creates an entity that is not contained in any of the combining parts. Emblems can represent signs of language, nominations, but not speech, for they are not able to convey the strict syntactic arrangements that are necessary for sentences.

3.1.4.   Emblems in Early Egyptian Script

We presuppose that emblematic type of writing was proper to many ancient systems at the very beginning of writing. The use of emblems can be seen, for example, in early Egyptian hieroglyphic inscriptions (Fig. 4), such as the Narmer Palette and the Scorpion Mace Head (both 32nd–31st c. BC), events are narrated through iconographic pictures while names are rendered phonetically. The name NARMER (Fig. 4a) (presumably for king Menes) *nˁr-mr* 'painful, stinging', or 'fierce catfish' is rendered as a combination of two glyphs. It is given three times: between the heads of cows (goddess Hathor) and near the king's head in the upper sector of the palette. There are other examples of small glyphs

near the people's heads, they should be their names. There is also a "number emblem" above the captive's head: 6 lotus flowers designate 6,000 captive warriors. The name of 'Scorpion' (Fig. 4b) is given in two images in front of the king's head.

(a)   (b)
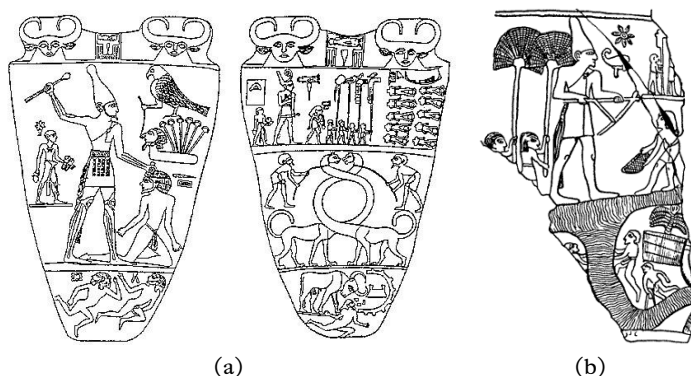
Figure 4. (a) The Narmer Palette (32nd–31st c. BC). (b) The Scorpion mace head (32nd–31st c. BC)

Examples of emblematic writing show that the formation of a linguistic emblem occurs at the very initial stages of writing, contributing to the allocation of simple pictorial components and their stabilization in the phonetic function.

3.1.5. Emblematic Techniques in Linear Scripts

The formation of linear writing is a gradual process, and it involves not only a fixed order of characters, but also the stabilization of their position and orientation on the line. In early examples of 'linear writing' (that is writing on a line, be it horizontal or vertical or something else), the sign can be rotated in different directions, i.e., it still exists as a pictorial image and not a written one. It acquires stable orientation when the line obtains a fixed one-dimensional orientation in the writing space.

Stable linear writing can also use the second dimension as additional space, combining linear elements in blocks or adding meaningful marks. This can be seen in hieroglyphic blocks of Mayan, Chinese, or Egyptian hieroglyphic, using the techniques of duplication and triplication of characters or combination of different characters in blocks—not only in their juxtapositions, but also including one in another; this is also attested in the Sumerian cuneiform system. It is also represented in abugida writing systems where the space around the invariant charac-

ter (akshara) allows the use of diacritics: superscripts, subsripts, post-scripts, prescripts, and even combinations of positions.

These positions can be used not only for vowel diacritics, but also for ligatures, subscript consonants, as well for pronunciation marks (such as nasalization). In Fig. 5a,b we can see an example: the well-known mantra 'Ō^m ma-ni pa-dme hū^m' in six syllable aksharas, in Devanagari and Tibetan scripts:

ॐ मणि पद्मे हूँ   ꊰ   옴마니반메훔

    (a)   (b)   (c)
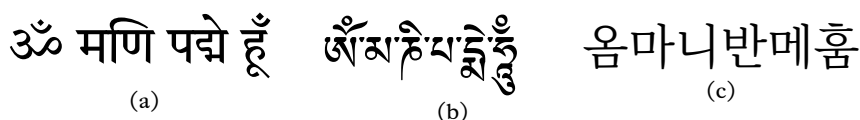
FIGURE 5. (a) Devanagari. (b) Tibetan. (c) Korean

The emblematic feature of 'new meaning creation' is proper to aksharas with vowel diacritics that acquire a new pronunciation status, which is not the sum of its components (/na/ + /i/ = /ni/, but not /nai/), and to ligatures: (/da/ + /ma/)+ /e/ = /dme/).

Block writing is also characteristic of Korean, where characters form syllable blocks that follow each other in linear order as in abugidas. In Fig. 5c we can see the same mantra written in the form of six Korean blocks.

The layout of characters in blocks allows the reading of components in a well-defined order; the enigmatic nature of emblem can be perceived only through distorted visual proportions of elements that make reading difficult to non-accustomed readers.

In many alphabetic systems, diacritic marks serve for differentiation of pronunciation.

Thus, in alphabets the characters with diacritics can acquire some properties of linguistic emblems: complex structure and new sound meaning.

It should be noted that the ligatured spellings can provide new characters which are then conceived as simple (not complex) signs. Among the examples we have <&> = Latin 'et', <?> that originates from the vertical arrangement of abbreviation "qo" of the Latin word "quæstio," <!> from Latin "Io," interjection of joy. But their emblematic nature goes forth in their special reference, in opposition with the surrounding context.

3.2.   Integral vs. Segmental Spelling

This dimension deals with the division of pronunciation units into parts in order to obtain a graphic representation. The starting point is the

word as an independent unit of speech. It may be rendered by an iconic pictogram, or by a pictogram of a homonym as a whole or by parts. This is a way to obtain a rebus representation and mixed phonetic spelling (like in Aztec emblems), as well as stable syllabic spelling.

A word (W) may be segmented differently, for example:

W = /CCVC/ = [CV]-[CVC] or [CV]-[CV]-[CV] or [CV]-[CV]-[VC], …
W = /CVCVC/ = [CV]-[CV]-[CV] or [CV]-[CVC], …

Natural segmentation gives a sequence of mora signs.

So we can have two types of spelling:

(1) using signs for close (and open) syllables (CVC, CCVC, CVCC, …)—lolo-type (*syllabic*)
(2) using signs only for moras (CV, V, -C) as minimal pronunciation units in decomposing a word—kana-type (*moraic*).

It is argued that not only Japanese refers to moraic systems, but also abugida and abjad refer to the same segmental class, since they have secondary ways for conveying long vowels into the syllable.

The decomposition trend seems to be opposite to the one of emblematic combination; yet it arises from the emblematic representation of complex units and contributes to the development of phonographic writing without support of meaningful units. Both trends coexist in ancient scripts.

3.3.   Complete vs. Reduced Spelling: Abugida and Abjad

Abugida with a standard subsystem of vowel modifications (C^V, V) is an example of complete vocalized writing: different vowels have different representations as independent signs or inside syllables.

Abjad is a graphically reduced, non-vocalized type, presupposing an indefinite vowel in a syllable (C^x), since a consonant cannot form a syllable per se, as pronunciation unit. The inherent vowel must be inferred from the context. So while abjad characters do not form 'emblems' as such, written words may have the property of only partial sound representation that requires the 'feeling of meaning' as do emblems.

Reduced spelling is proper to Egyptian hieroglyphic writing with uniliterals C^x, biliterals $C^x\, C^x$ or triliterals $C^x\, C^x\, C^x$.

Reduced systems are also consonantic alphabets, where vowels have stable diacritic forms (not always used).

In all these non-vocalized systems, a vowel is conceived as the inherent characteristic of a syllable (mora), is variable in word formation and cannot begin a syllable.

Examples of reduced spelling of another nature can be also found in Aztec emblems, where the locative suffixes of place names are often

omitted. The logographic *dongba* script is reduced, due to the absence of grammatical indexes; a similar case is probably the one of the *rongorongo* script of Easter Island (not yet deciphered).

Both trends serve to support the interests of users: complete spelling serves to readers, and reduced spelling, to writers.

3.4.   Simple vs. Differentiated Spelling

Differentiation is an important device in the development of writing. It includes different techniques. Semantic differentiation appears in the use of graphic determiners in ancient writing systems. These are pictograms with classifying function, so they are hyperonyms to the determined units and serve to differentiate homonyms. Another sort of semantic differentiation in logographic writing, given by H. Rogers, is the use of additional minor graphic signs for specialization of meaning; it occurs in Sumerian writing, when small cuneiform strokes are drawn inside the logogram HEAD as indication for teeth, in order to express the meaning of 'mouth' (Rogers, 2005, pp. 88–89).

Phonetic differentiation presupposes the use of elements that refine the reading of a simple sign; it works already on the morphosyllabic level (as phonetic complementation), for example in Egyptian script:

$(n^x f^x r^x) + f^x + r^x$ = /nefer/ 'beautiful' .

Abugida differs from kana systems using diacritic vowel modifications of the invariant sign, whereas kana uses several invariant signs for different vocalization (Fig. 7b). Yet kana differentiates diacritics for voiced and unvoiced pairs (Fig. 7c)
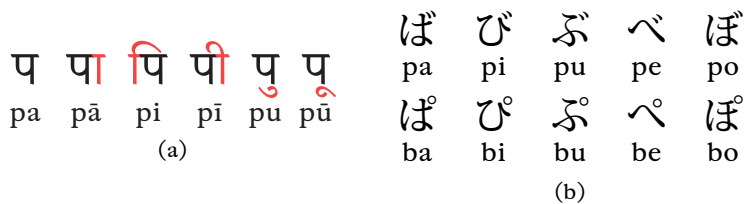
Figure 6. (a) Devanagari. (b) Japanese kana

Alphabetic writing is the last stage of phonological analysis.

According to alphabetic principle, every phoneme, consonant or vowel must be represented by a full-formed grapheme.

Abjad writing is largely defined by the phonological, morphological, and lexical structure of classical West Semitic languages, where a vowel

is not an independent unit: it cannot start a syllable, and it is a word-formation variable (and not a constant attribute of the root).

Alphabetic writing appears in languages in which vowels have independent values, so that they are represented by characters equivalent in size and position to consonant letters.

We can allow the metaphor of democracy here (with 'gender' sense): vowels are hidden under yashmak in the presence of consonants in abjad, they form different "garments" for consonants in abugida (sometimes they form the 'soul' of a consonant 'body' in an akshara), and, finally, the Greek claim for democracy gives them their independent status in alphabetic text.

4.   Some Concluding Comments

Thus, the main points made in this work concern:

(1) the role of linguistic emblems in the formation of phonographic writing;
(2) the representation of the evolution of writing based on the psychologically distinguishable units of speech and language, which are fixed in written signs: the word as a semantic and phonetic unit and the syllable (mora) as a pure pronunciation unit;
(3) the alleged moraic nature of abjad and abugida writing;
(4) the use of binary classification features (linear/nonlinear, integral/decomposed, complete/reduced, simple/differentiated spelling) for developed typological schemes of writing systems.

Thanks to the four characteristics of writing systems mentioned above, we can describe transfers from one type to another. Thus, abjad differentiated by diacritics becomes a consonant-alphabet. Alphabets using techniques of non-linear block spelling may be designated as a separate type. The next step of differentiation deals with featural Korean script.

The proposed scheme can be further detailed; some additional classification criteria, taking part in the way a writing system is functioning, can be identified as follows:

- amount of ideography (not only of morphography),
- amount of xenography (taking into account the use of graphemes of foreign, different languages, xenograms, or heterograms),
- level of graphic complexity (analytic/synthetic writing, the latter presupposing the use of complex graphemes, cf. Fedorova, 2012),
- orthographic depth (according to Rodgers),
- level of semiotic heterogeneity (with respect not only to different languages, but also to graphic systems of different semiotic nature, cf. Perri, 2014).

These dimensions of writing need special investigation. Only a few comments can be made. Thus, modern devices allow using different sorts of icons along with written words; this is a case of mixed writing. The use of emoticons becomes common for informal communication all over the world, for they give expressive images of emotions, as in these Japanese examples:

(˄˄)!   (*0*)   \(˄_˄)/

Another case of mixed writing happens in the simultaneous use of Latin and Cyrillic (or another national) writing signs; it is often a result of contacts of languages and of their writing systems (ibid.).

Let us also mention a particular language game, using number codes. This method occurs widely in advertisements and informal Internet communication. Here are examples of this kind of "numeric codification," in Chinese phrases:

886 /bā bā liù/ = 拜拜了 /bàibài le/ 'Bye-bye'
768 /qī liù bā/ = 吃了吧 /chī le ba/ 'Let's go eat'!

Multilingual and multiscript texts on bill-boards are common practice in modern cities, forming their linguistic landscape.

The contrast between the writer's and the reader's interest, contribute to the development of writing. It may not be so much about evolution as about writing improvement. Different forms of writing co-exist in the modern world, addressing different needs: speed, exactness of speech transfer, the best visual presentation of content or just of its form… Writing can serve not only for distributing information, but also to conceal it; it can be a means of magic, or play, of expressiveness, or of decoration. But all of these mixed forms and techniques can exist only as deviations of existing standard writing systems or as graphic games taking advantage of the creative potential of the art of writing.

References

Berdan, Frances F. (1997). "The Place-Name, Personal Name, and Title Glyphs of the Codex Mendoza: Translations and Comments." In: *The Essential Codex Mendoza*. Ed. by Berdan Frances F. and Patricia Rieff Anawalt. Vol. 1. Berkeley: University of California Press, pp. 163–239.

Berdan, Frances F. and Patricia Rieff Anawalt (1997). "Glyphic Conventions of the Codex Mendoza." In: *The Essential Codex Mendoza*. Vol. 1. Berkeley: University of California Press, pp. 93–102.

Berlin, Heinrich (1958). "El glifo "emblema" en las inscripciones mayas." In: *Journal de la Société des Américanistes* 47, pp. 111–119.

Bradley, David (2009). "Language policy for China's minotities: Orthgraphy development for the Yi." In: *Written Language & Literacy* 12.2, pp. 170–187.

Bright, William (2000). "A matter of typology: Alphasyllabaries and Abugidas." In: *Study in the Linguistic Sciences* 30.1, pp. 63–71.

Daniels, Peter T. (2009a). "The study of writing systems." In: *The World's Writing Systems*. Ed. by P. Daniels and W. Bright. New York: Oxford University Press, pp. 3–17.

————— (2009b). "Two notes on terminology." In: *Written Language & Literacy* 12.2, pp. 258–274.

Dyakonov, Igor M. [Дьяконов, Игорь М.] (1976). "Протошумерские иероглифы [Proto-Sumerian hieroglyphs]." In: *Тайны древних письмен: Проблемы дешифровки [The mysteries of ancient scripts: Problems of deciphering]*. Ed. by Igor M. Dyakonov [Игорь М. Дьяконов]. Москва [Moscow]: Прогресс [Progress], pp. 569–571.

Elkins, James (2003). "Four ways of measuring the distance between alchemy and contemporary art." In: *HYLE—International Journal in Philosophy and Chemistry* 9, pp. 105–118.

Fedorova, Liudmila (2009). "The Emblematic Script of the Aztec Codices as a Particular Semiotic Type of Writing System." In: *Written Language & Literacy* 12.2, pp. 258–274.

————— (2012). "The development of structural characteristics of Brahmi script in derivative writing systems." In: *Written Language & Literacy* 15.1, pp. 1–25.

Gelb, Ignace J. (1963). *A study of Writing*. Chicago: Chicago University Press.

Istrin, Viktor A. [Истрин, Виктор А.] (1965). *Возникновение и развитие письма [The appearance and development of writing]*. Москва [Moscow]: Наука [Nauka].

Lee, Sang-Oak (2009). "The Korean alphabet: An optimal featural system with graphical ingenuity." In: *Written Language & Literacy* 12.2, pp. 202–212.

Perri, Antonio (2014). "Why writing is not (only) transcribing? Writing codes in contact: steps towards multigraphic literacy practices." In: *Testo e Senso* 15, pp. 75–98.

Rogers, Henry (2005). *Writing systems: A linguistic approach*. Malden, MA and Oxford, UK: Blackwell Publishing.

Siméon, Rémi (1963). *Dictionnaire de la langue nahuatl ou mexicaine*. Ed. by Alex Wimmer. Graz, Austria: Akademische Druck- u. Verlagsanstalt.

Sullivan, Thelma D. (1983). *Compendio de la Gramática Nahuátl*. México: Universidad National Autónoma de México.

The Naasioi Otomaung Alphabet of Bougainville

A Preliminary Sketch From Afar

Piers Kelly

Abstract. The Naasioi Otomaung alphabet first came to light during the Bougainville Crisis of 1988–1998. Created by the Naasioi-speaking leader of a politico-religious movement in Kieta district, its emergence follows the pattern of numerous other scripts of Asia and the Pacific that have developed in recent times in the context of anti-colonial confrontations (Kelly, 2016; 2018a). This paper provides the first ever public report on the form, structure and context of the script, early efforts at documentation, and its prospects for future development. The script exhibits a formal influence from cursivised Roman while its inventory of letters presents as a cypher for the English alphabet, including letters such as <x> and <z> that are not present in standard Naasioi orthographies (Hurd and Hurd, 1966). From the perspective of its users, however, the alphabet is designed to universally encode any language: the word *otomaung* is in fact a neologism roughly meaning 'able to express anything'. The term is also polysemous, variously denoting the letter <A>, as well as the religious community in which the alphabet was created. The forms of the letters, meanwhile, are said to have been inspired by ceremonial scarring, a practice that is now rare. Reproducing these forms in writing is thus seen as an act of cultural preservation by other means. Although at one time the script became part of a local school curriculum, literacy is now limited to a small number of individuals. Systematic documentation and description of Naasioi Otomaung has suffered various setbacks, from political disruptions to the COVID-19 pandemic. As a result, most of the documentation to date has been carried out by post, email, and social media correspondence. Despite the obvious limitations and inefficiencies of these channels, 'virtual' fieldwork has been unexpectedly productive, resulting in an accurate record of the script, preliminary information about its historical and ethnographic circumstances and the development of a new font. With Bougainville's recent advances towards political independence, the Otomaung Naasioi alphabet may soon rise to greater prominence.

In this paper I partially describe the Naasioi Otomaung, a recently devised script of the Autonomous Region of Bougainville, Papua New Guinea. In broad terms I outline its formal and typological features as

Piers Kelly   🆔 0000–0002–6467–2338
Department of Archaeology, Classics and History Rm 308, C02 Building
University of New England, Armidale NSW 2351, Australia
E-mail: pkelly26@une.edu.au

well as its history and ethnographic context. On account of local polit-
ical disruptions and the COVID-19 pandemic, face-to-face fieldwork in
the Naasioi-speaking region of Bougainville has not yet been feasible.
Nonetheless, with the aid of mobile phone calls, email and especially
social media I have recorded foundational information on the history,
ethnographic context, formal properties, structure and uses of this new
script. Thus, in addition to providing primary documentation and de-
scription, this paper is intended to demonstrate that digitally mediated
fieldwork can produce surprisingly rich results. In turn, it is my hope
that this preliminary work on the Naasioi Otomaung alphabet will serve
as a secure basis for future ethnography based on face-to-face partici-
pant observation in the field.

Motivation

Primary writing systems, and their derivatives, have always been a
major focus of attention for palaeographers and grapholinguists. The
analysis of these systems has generated insight into the origins and evo-
lution of writing, permitted the diachronic reconstruction of script lin-
eages, and set the parameters for establishing grapholinguistic typolo-
gies. Secondary scripts that have been deliberately devised in recent
times have received less attention, perhaps because they do not seem to
offer any imposing insights into the nature of writing. I have argued,
however, that there is much to learn from secondary scripts, especially
those that have been invented within small-scale, non-state societies in
the context of recent colonial contact (Kelly, 2018a,b). The Cherokee
and Vai scripts are well-known examples of this phenomenon but many
others have also been documented and analysed.

Like all good objects of anthropological enquiry secondary scripts
are self-evidently diverse while having surprising features in common.
This diversity-universality axis can be approached in a straightforward
grapholinguistic mode that attempts to describe and compare formal
and systemic characteristics. However, by adopting wider perspective
afforded by the anthropology of literacy paradigm, we can also attend
critically to the historical determinants and political contexts of these
scripts, as well as the various functions that they serve, and the cultural
meanings that their users ascribe to them.

My research to date has focused on West Africa and Southeast Asia,
two regions in which a bewildering array of new scripts have been in-
vented within non-state societies over the past two centuries. Certain
scripts such as the N'ko script of the Côte d'Ivoire and the Kayah Li
script of Thailand-Myanmar have large communities of contemporary
users. Others such as the Bagam script of Cameroon and the Sulit Air
script of Indonesia are known from only a few fragmentary manuscripts.
Others still, including Pa Chay script of Vietnam and Pahawh Khmu'

script of Laos have no surviving inscriptions and are recalled only in oral histories. Individual inventors in these two regions continue to develop secondary scripts, while every year scholars unearth more that have been overlooked in informal archives. Needless to say, the documentation of secondary scripts is far from complete.

Despite the growth of anthropological scholarship in this arena, with the important work of Cécile Guillaume-Pey, Konrad Tuchscherer, Carmen Brandt and others, I have often encountered resistance to their study. In conversation, some colleagues have expressed the view that they are not 'real' scripts, and that there are endangered scripts and languages that deserve more documentary attention from researchers. Others have pointed out that recent secondary scripts are rarely successful, especially if success is measured by the extent of the diffusion of the script and its transmission over multiple generations. Another objection is that the scripts themselves are often structurally cumbersome and that they simply add a distraction to the more important goals of orthography development for minority languages, and ultimately literacy in these languages.

On the whole, these objections are premised on utilitarian concerns and on implicit hierarchies of value where the relevant parameters of interest are the age of the scripts, their degree of 'naturalness' or 'authenticity', and their structural efficiency vis-à-vis the languages they are intended to represent. If such values and concerns are universally held, then we can register these objections as perfectly legitimate. However, numerous studies of literacy ideologies, of which Brian Street's contribution (Street, 1984) is most well recognised, demonstrate that such values are very much culturally and historically positioned and require ethnographic explanation in their own right.

From my perspective, new scripts deserve grapholinguistic and anthropological attention for the additional reason that they offer a rare insight into the diversity of human symbolic culture. At the same time, they can help us perceive what is *un*diverse about the way we do things with graphic codes in terms of both the 'obvious' and non-obvious solutions and processes that we often converge upon to address common problems. Moreover, small-scale or non-state societies are ideal sites for investigating written practice on the basis of the fact that they represent locations where writing, of any kind, has been a relatively recent introduction. Consequently, normative literacy attitudes have not yet had a chance to become hegemonic in the same way that they have in the West, where it is no longer possible to participate fully in society unless you are literate.

Diversity and Universality of Secondary Scripts

Three brief examples suffice to provide a glimpse into the diversity of new scripts. The Bamum script of Cameroon (developed between 1896

and 1910) includes a semantic determinative for distinguishing homophones (Dugast and Jeffreys, 1950). The homophonous lexeme that is marked with the determinative is the one with the meaning that seen to carry more prestige (Fig. 1).

FIGURE 1. Homophonous terms in the Bamum language distinguished in writing with a semantic determinative (Dugast and Jeffreys, 1950, p. 7)

The Western Apache script was designed as a prompt for the recitation of prayers and it includes signs designed to specify the correct ritual gestures that accompany the speech (Fig. 2). These so-called kinetic signs are generated as compounds of the speech signs meaning that it requires deep insider knowledge to be able to read and reproduce the script.

FIGURE 2. Western Apache script (Basso and Anderson, 1977, p. 232)

Finally, the Sayaboury script of Laos includes signs for signalling vocal noises such as chanting and calling animals (Fig. 3).

These three examples draw attention to the fact that culturally specific, non-phonographic and even non-linguistic information can be encoded in graphic form, a fact which presupposes that in order to learn and use the script effectively one must also be a competent participant in that society. These examples are not provided merely to draw attention to interesting or quirky outliers. Rather, they are sharp illustrations of the fact that writing of any kind is culturally loaded, and that it

FIGURE 3. Sample of the Sayaboury script (Smalley and Wimuttikosol, 1998, p. 115)

is never a neutral or autonomous mechanism for representing language (Bartlett, Lopéz, Vasudevan, and Warriner, 2011). It is by stepping outside of our own literacy context we can acquire a better appreciation for the inherent relativity of scripts.

Having made the case for relativism it is also possible to make productive generalisations about the ways that secondary writing systems emerge and evolve. New scripts, especially those invented by non-literates, seem to exhibit high visual complexity or iconicity, have no contrast in reverse or rotated images, and to become compressed over multiple transmissions. They are also often morpheme-centric with a preference for representing syllables and consonants as opposed to individual vowels, and to make use of rebuses and semantic determinatives. More research is needed in order to ascertain their convergent techniques for modelling language, and the extent to which their dynamics may coincide with those of primary scripts.

The Linguistic Context of the Naasioi Otomaung Alphabet

A very recent secondary script that has not yet been formally documented is the Naasioi Otomaung alphabet from the island Bougainville. I was first told of the existence of this alphabet by the Bougainvillean linguist Ruth Spriggs, but have never had the opportunity to investigate it in person. The COVID-19 pandemic has been the most formidable obstacle to research, but it also had the effect of liberating me from any expectation—and guilt—that face-to-face fieldwork was at all possible. My preliminary documentation, offered here, is the result of mobile phone calls, social media interactions and generous work performed at my direction by intermediaries already on the ground including missionaries and linguists.

Despite its small geographic size, Bougainville is very linguistically and culturally diverse. The coastal languages marked with a star on the map below are Austronesian and the mostly inland languages marked in

grey are Papuan. This is a contrast that still reflects the earlier coloni-
sation of Bougainville by Austronesians some three thousand years ago.
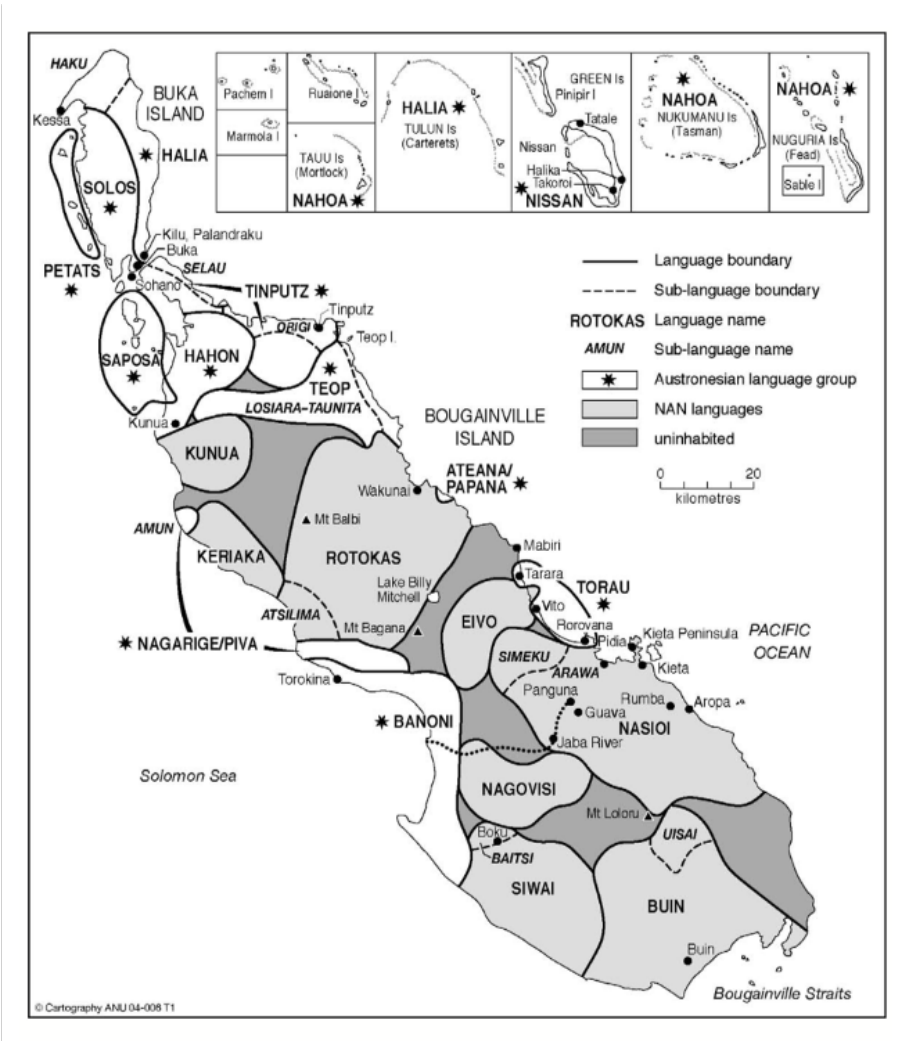
FIGURE 4. The languages of Bougainville (Tryon, 2015, p. 32)

Rotokas, a Papuan language of Bougainville spoken by an estimated
4,320 people, is famous for having what is claimed to be the smallest
sound inventory in the world, with only 11 contrastive phonemes. The
efficiency of the Rotokas sound system has even inspired the invention

of a so-called 'polysyllabary' on the part of linguist Sheldon Ebbeler
(Ebbeler, 2014), though this has not been adopted by Rotokas speak-
ers. Meanwhile, the Naasioi language, spoken further south, has about
20,000 speakers and has an inventory of 19 phonemes. It was a speaker
of Naasioi who was responsible for creating the Naasioi Otomaung al-
phabet described in this paper. A representation of the Naasioi phoneme
inventory alongside the Naasioi Otomaung alphabetic signs that are as-
signed to its sounds, is provided in Fig. 5 below.

FIGURE 5. The phoneme inventory of Naasioi with Naasioi Otomaung letters

As the above chart indicates the Naasioi Otomaung script does not
distinguish vowel contrasts, and the engma sound is rendered with a
digraph. Of further interest is the fact that there is no glottal sign despite
the relatively high functional load of glottal stops in Naasioi.

These curiosities can be explained with reference to the historical
context of the system. The Naasioi Otomaung alphabet was created
by a man known as Chief Peter Karatapi who is also credited with the
founding of the Otomaung cultural movement, from which the alpha-
bet emerged. The alphabet enjoys popularity among Naasioi speakers
living on the Siang river, but only a few profess any literacy in the sys-
tem. I know of only three fully literate individuals by name: Chief Peter
Karatapi, his daughter Maryanne Karatapi and Steven Tamiung. There
are probably many more who are partially literate in it. There are not
any taboos or restrictions against learning or disseminating the script,

and it's users are hoping to promote a greater role for it in a future independent Bougainville.

Form and Function

Naasioi Otomaung is a straightforward cypher script for the Roman alphabet. I characterise the difference between an ordinary writing system and a cypher script, in the following way: a writing system is designed to model aspects of linguistic structure, usually phonological, of a language or languages. A cypher script, meanwhile, models another writing system. In other words, it is a graphic representation but at one remove and could be thought of as a form of radical typographic differentiation. Cypher scripts can have different functions and motivations. Broadly, they can be used to make a writing system readable in another modality, for example in Morse code. Equally they can be used as a kind of graphic encryption or disguise, or they can be designed to do the political work of projecting an ethnolinguistic contrast.

In Naasioi Otomaung there are three typographic 'registers' that coincide with uppercase, lowercase and a third ornamental register which is perhaps like bold. The tabulation in Fig. 7 below has been provided by Steven Tamiung.

The signs lavelled as 'Handwriting Economy,' are repeated to the bottom of the grid.

Fig. 8 is my own reorganisation of this chart. What Fig. 8 illustrates is that Naasioi Otomaung models the 26 letters of the English Roman alphabet, and includes letter signs for sounds like x and z that aren't even attested in standard orthographies of Naasioi. Among other things, this is because Naasioi Otomaung is also used for writing English and Tok Pisin. The table also reveals that, to generate the ornamental register, the writer takes a lowercase sign then adds a superscript feature in the form of a horizontal bar and series of five dots, sometimes rendered as small vertical lines, then an ornamentation on top that has no linguistic or semantic content, as shown in Fig. 9 below.

What is also evident from Figure 8 is that Naasioi Otomaung is not just a Roman cypher at the level of an alphabetic system, but there is also a Roman influence in the morphology of the script. There is stereotyping in its consistent slant and the use of ascenders. The letters indexOtomaungOtomaung k (<k>) and ⟨ (<n>) are distinguished on the basis of their orientation, like a <b> and <d> contrast, but other letters appear to be distinguished by means of very subtle graphic elements. It is also clear that the recitation order, which is the English recitation order, has influenced the design, meaning that letters that are adjacent in the recitation sequence are often graphically similar and are

FIGURE 6. Chief Peter Karatapi at the 2019 Bougainvillean referendum. Image: Steven Tamiung

FIGURE 7. Tabulation of the three registers of of Naasioi Otomaung

FIGURE 8. Table of Naasioi Otomaung signs

Upper	Lower	Ornamental

FIGURE 9. Rule for generating signs in the ornamental register

distinguished by means of one very small visual contrast. This situation brings about challenges for legibility. No substantial manuscripts or inscriptions in Naasioi Otomaung have been made available to me, however the following example sentence was produced by Maryanne Karatapi using upper and lowercase letters, and has been interlinearised (1) by Jason Brown:

(1)   *Mosi kuaparamang.*
       ⟨script⟩
       Mosi   kua-p-a-ra-mang
       Dog   run-3-REFL.SG.PRS.PRG
       'The dog is running'

Just before this paper went to press, Steven Tamiung sent a chart of Naasioi Otomaung numerals, shown in Fig. 10 below.

More analysis of numerals is needed though it is interesting to note the similarity with the Hindu-Arabic numeral system in signs for 0 and 1, and the apparent use of tally marks in the signs for 2 3, 5, 6, 7 and 8. The core forms of 1, 4, and 6 are also recycled in subsequent numerals, a process also observed in the numeral system of a new script invented by a nativist movement of the Philippines (Kelly, 2016).

Historical Background

Naasioi Otomaung was created at the very epicentre of what has come to be known as the Bougainville Crisis (1988–1998), a ten-year violent conflict that can be characterised as both a struggle for political independence from Papua New Guinea and a struggle for political unity within

Figure 10. Naasioi Otomaung numeral set

Bougainville itself. The island has experienced successive waves of colonisation. From the 1880s it was an imperial possession of Germany, and after WWI it became, along with Papua New Guinea, a colonial possession of Australia. During WWII it was occupied by Japan and America, and was eventually returned to Australia in 1946. Naasioi-based movements opposing Australian rule began in the 1960s and shortly after, the controversial Panguna copper mine was established by a subsidiary of Rio Tinto in Naasioi country in defiance of local opposition. The mine caused devastating environmental damage and exacerbated existing secessionist agitation throughout the 1970s. It was in 1975 that Papua New Guinea was granted independence from Australia, meaning that the new PNG government took over the administration of Bougainville and supervision of its mine.

The Panguna mine continued operation throughout this period and by 1988 there was outright war. The principal military actors in this conflict were the Bougainville Revolutionary Army or BRA and the PNG defence force. Hostilities did not come to an end until 1998 and a peace agreement was eventually signed in 2001. In late 2019, the PNG government held a non-binding referendum, in which the overwhelming majority of voters opted for full independence.

Although my characterisation of events is a reduction of highly complex situation with many variables, a decisive aspect of the war was the long-term blockade that the Papua New Guinea government placed on Bougainville from 1990 to 1994. In this time no people or goods were permitted to enter or leave the island in a strategy that was intended to weaken the BRA and force its surrender. The blockade pro-

duced enormous hardship, but it also became a catalyst for extraordinary innovation to ensure survival and self-sufficiency. Among other initiatives, local communities repurposed abandoned mine equipment to create home-made hydroelectric power plants, and produced their own biodiesel from coconuts to keep vehicles running. These technological innovations and initiatives reinforced the idea that Bougainville was quite capable of autonomy and that genuine independence was within reach. Memories of the blockade are an important historical reference point for Bougainvilleans today, especially in the context of the COVID-19 pandemic (Fig. 11).

FIGURE 11. Facebook status update from Dickson Marcelline Karatapi, 6 June 2020

Before and during the conflict Naasioi people were known to join various new cultural, religious and political movements, of which the Bougainville Revolutionary Army was just one example. Across the island, competing micronationalist movements emerged that replicated all the structures of nation states, sometimes with banks, police forces and civil administrations including parliaments. The Otomaung cultural organisation that produced the Naasioi Otomaung alphabet, was a nativist movement concerned with cultural revitalisation. Its agenda was to restore, preserve and promote indigenous cultural forms including rites, songs, ceremonies and dances.

The Otomaung movement continues to be led by its founder, Chief Peter Karatapi, whose ambitions once included the establishment of a culturally authentic and independent education system for Naasioi speakers. His indigenous schools replicated traditional subject areas of the PNG education system but replaced the content with native alternatives: Naasioi language was taught instead of English, traditional religion replaced Catholicism, while literacy instruction took place in the Naasioi Otomaung alphabet. At that time, Karatapi referred to the alphabet as *Me'ekamui Kepia*, with the ascribed meaning of 'Bougainvillian alphabet'. The choice of the term *Me'ekamui* (which can be translated as 'holy island' or 'sacred place') points to a likely influence from his associate Damien Dameng, the founder of a radical secessionist organisa-

tion known as *Me'ekamui Onoring Pontoku,* roughly meaning "government of the guardians of the sacred land" (Regan, 2002). Dameng rejected all foreign influences and the three precepts of his movement were that "Western education belongs to the bad spirits; Western health belongs to the dogs; and Western religion belongs to immature kids" (Roka, 2014). In the 1990s, Dameng's movement became part of the ideological inspiration for the Bougainville Revolutionary Army under its leader Francis Ona (Hermkens, 2013). It has even been argued that Francis Ona adopted Dameng's program in order to shore up waning political support.

The Otomaung cultural movement, Me'ekamui Onoring Pontoku, and the Bougainville Revolutionary Army all coexisted in central Bougainville during the conflict and they probably had overlapping memberships to a degree, but towards the end of the fighting, the BRA denounced Otomaung as a cult and began persecuting its members until a peace was established between the two groups in June of 1997 (James Tanis, pers. comm.). The overtly pacifist philosophy of the Otomaung movement was no doubt fundamentally at odds with the recruitment aims of the BRA.

Members of the Otomaung movement were later invited to perform at the signing of the Bougainville Peace Agreement on 30 August 2001.

Literacy Practice, Uses and Meanings

During the Bougainville Crisis (1988–1998), the Naasioi Otomaung script was used in designs on clothing and the missionary linguists Conrad and Phyllis Hurd recall seeing it embroidered into a dancing cape. For a short while it entered Peter Karatapi's alternative school curriculum in Kieta district where it was taught up until third grade. I do not presently have a clear view of how the script is actually used today, beyond inscriptions on objects including t-shirts (Fig. 12), fans (Fig. 13) and political banners (Fig. 14). It is also used for the informal teaching of those who want to learn it, as well as in demonstrations to outsiders like me. However there are three distinct aspirational uses for the script that I have identified from my direct and indirect discussions with practitioners who promote it as a universal writing system, as a mechanism for preserving cultural knowledge, and as a visible embodiment of indigenous cultural values.

Universal Writing System

The word Otomaung, discussed further below, is a neologism with the ascribed meaning of 'able to express anything'. Consistent with this

FIGURE 12. A Naasioi Otomaung in-
scription on a t-shirt. Image: Steven
Tamiung
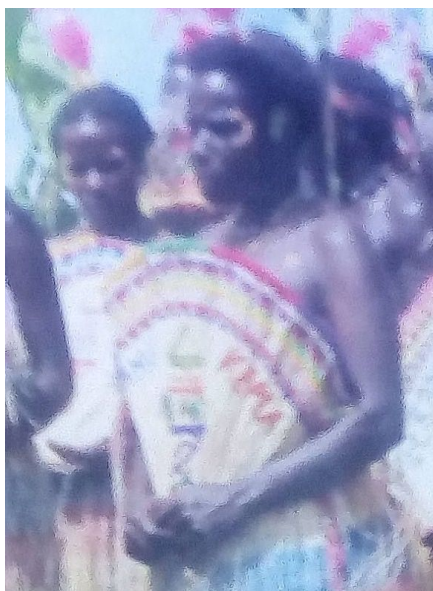
FIGURE 13. Inscription in Naasioi
Otomaung woven into a fan held by a
Naasioi performer in Buka, 2001. Im-
age: Steven Tamiung.

FIGURE 14. Naasioi Otomaung peace banner. From left to right: Pionu
Pantadera, Nekenung Butung, Cecilia Nenominu, Alice Piokanu and Theresa
Bangsingona.

meaning the Naasioi Otomaung script is promoted as a kind of utopian universal system for representing all the languages of Bougainville, including English. In fact the writing system itself was simply one innovation in a micronationalist package of replacement or parallel systems that include an indigenous currency, an indigenous lunar calendar and the promotion of Naasioi as a national auxiliary language to unite Bougainville. I have not had access to information about the lunar calendar or currency but it bears pointing out that another movement, still active in southern Bougainville has created its own micronation with patrolled borders and which does in fact have a separate currency (Cox, 2013).

Preservation of Cultural Knowledge

A second aspirational use of the Naasioi Otomaung script, according to Peter Karatapi, is to record and preserve information of cultural value including songs, dances and stories. The chart in Figure 15 below, for example, is apparently representing a musical stave, with each note of an octave marked in the rows. More research is needed to explain how the chart is to be interpreted musically, and why the ornamental sign for <m> is repeated within the stave.

   Beyond this one example I'm not personally aware of the existence of a manuscript tradition that is in fact recording traditional knowledge. But whether or not such an archive exists, the script itself is seen to index indigenous culture in its form.

The Embodiment of Indigenous Values

This leads directly to the third aspiration for the script: that it is capable of embodying indigenous cultural values in its graphic morphology.

   Before World War II, Bougainvilleans of various language groups were known to engage in practices of ceremonial body scarring or cicatrisation. Steven Tamiung, told me that oldest daughter in a Naasioi family usually underwent ceremonial scarring on her thighs at first menstruation but this was no longer performed. For the promoters of the Naasioi Otomaung alphabet, the script is seen to represent these once-prevalent sacred designs. The ethnographic record, though sparse, indicates that cicatrisation was not limited to women and continued to practiced well after the war among the Naasioi as well as other groups (Emanuel and Biddulph, 1969). Few analyses of cicatrisation on Bougainville have ever been published. The earliest known to me is the brief account of Naasioi scarring provided by Ernst Frizzi-München (1914). More detailed is the ethnography of Beatrice Blackwood (1935), centering on Kurtatchi village in the Tinputz-speaking region of Bougainville.

FIGURE 15. Musical stave rendered in Naasioi Otomaung

Figures 16 and 17 below are derived from these works. I have traced the cicatrisation patterns in turquoise in order to increase their visibility.

These patterns impressionistically display stylistic similarities with certain signs in the Naasioi Otomaung ornamental register, specifically in the arrangements of geometric lines and dots (Fig. 18). Here I do not wish to make any strong claim that the Naasioi Otomaung script is iconic of ceremonial scarring or that it demonstrates a direct cultural continuity with these practices. Nonetheless this is a value expressed by users of the script.

Another way in which Otomaung embodies cultural values is in the recitation names of individual letters of the alphabet some of which are supposed to be derived from the Lord's Prayer in Naasioi, and they each index a particular value.

The letter <A>, for example is named *otomaung*, and it gives its name to both the alphabet and the movement. We can see this word in the first line of the Lord's Prayer in the two translations of it that have been made available to me in Naasioi :

Niuma da <u>otomaung</u> pangningko, miring dakanaa mmeka'angta angpinang pangningkong pi'na.

Niuma paning-koo <u>otomaung</u>, dakaang miring meeka'antavari otoaing.

'Our father in heaven, hallowed be your name'
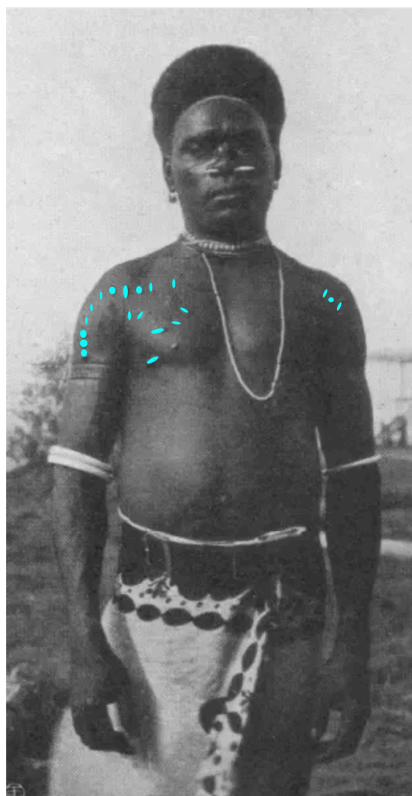
FIGURE 16. Traditional Naasioi scarring pattern (Frizzi-München, 1914, p. 44)

FIGURE 17. Scarring patterns in Kurtatchi village (Blackwood, 1935, p. 430)

FIGURE 18. Two standard Kurtatchi patterns identified and sketched by Black-wood (1935, p. 431), above. Below them are the ornamental signs for <s> (left) and <r> (right).

The letter <b> is named *miru* and this too is supposed to be a de-rivation from the Lord's Prayer, but the closest I can discover is *miring,* above.

Meanwhile, the final four letters are:

W = Siouma
X = Nari
Y = Kapoo
Z = Tampara

Together they form the sentence 'Siouma Nari Kapoo Tampara' which Tamiung describes as an expression of peace in Naasioi. I haven't been able to interlinearise this in Naasioi, but Tamiung has provided the following pragmatic gloss "1. come to the roundtable discussion to solve problems; 2. do not take law in your own hands; 3. Solve differences in words rather then actions."

Just before this paper went to press, Tamiung sent me a chart of Naa-sioi Otomaung signs with associated meanings (Fig. 19). These signs are not part of the alphabetic set and could thus be provisionally analysed as logographs.

Tamiung explained that the values encoded in these signs were taught as the 'twelve principles' that students were required to learn as part of religious studies in traditional school system established by Peter Karatapi.
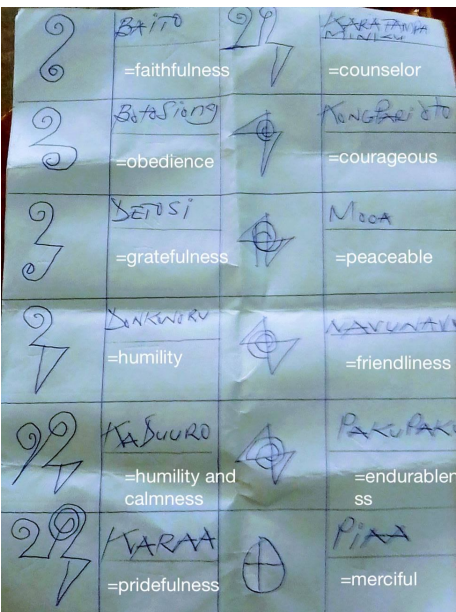
FIGURE 19. Naasioi Otomaung logographs

Summary

The Naasioi Otomaung alphabet was created by cultural leader Chief Peter Karatapi in the context of the Bougainville Crisis (1988–1998). The script was devised as part of broader cultural package within in a project of political and cultural autonomy. Its three utopian goals are to serve as a universal script for Bougainville, to be a method for recording and preserving cultural knowledge, and to embody indigenous cultural values.

The system itself is a cypher for the Roman alphabet, and includes a digital numeral system and a set of at least 12 logographs. Individual alphabetic letters are also invested with semantic values in some contexts. In practice it has been taught in a traditional and autonomous school system, and has been used for inscriptions on objects including traditional fans, dancing capes, T-shirts and banners. Many more questions need to be addressed regarding its history and meaning, the extent of literacy in the script and its prospects for the future. In the meantime, I hope that this paper has revealed the value of distance fieldwork, and demonstrated just how much information can be assembled from afar when travel is not feasible.

Acknowledgments

References

Bartlett, Lesley et al. (2011). "The anthropology of literacy." In: *A companion to the anthropology of education*. Ed. by Bradley A. Levinson and Mica Pollock. John Wiley & Sons, pp. 154–176.

Basso, Keith H. and Ned Anderson (1977). "A western Apache writing system: The symbols of Silas John." In: *Sociocultural Dimensions of Language Change*. Elsevier, pp. 227–252.

Blackwood, Beatrice (1935). *Both sides of Buka passage: An ethnographic study of social, sexual, and economic questions in the North-Western Solomon Islands*. Oxford: Clarendon Press.

Cox, John (2013). "The magic of money and the magic of the state: Fast money schemes in Papua New Guinea." In: *Oceania* 3, pp. 175–191.

Dugast, Idelette and M.David W. Jeffreys (1950). *L'écriture des Bamum: Sa naissance, son évolution, sa valeur phonétique, son utilisation*. Vol. 4. Populations. Paris: Mémoires de l'Institut Français d'Afrique Noire.

Ebbeler, Sheldon (2014). "Uriovakiro: a polysyllabary for Rotokas." https://www.omniglot.com/pdfs/uriovakiro.pdf.

Emanuel, Irvin and John Biddulph (1969). "Pediatric field survey of the Nasioi and Kwaio of the Solomon Islands." In: *Journal of Tropical Pediatrics* 15.2, pp. 55–69.

Frizzi-München, Ernst (1914). *Ein Beitrag zur Ethnologie von Bougainville und Buka, mit spezieller Berücksichtung der Nasioi*. Leipzig & Berlin: B. G. Teubner.

Hermkens, Anna-Karina (2013). "Like Moses who led is people to the Promised Land: Nation and state-building in Bougainville." In: *Oceania* 3, pp. 192–207.

---

1. https://bravenewwords.info/ottomaung-font-project/

Hurd, Conrad and Phyllis Hurd (1966). *Nasioi language course*. Port Moresby: Summer Institute of Linguistics.

Kelly, Piers (2016). "Introducing the Eskaya writing system: A complex Messianic script from the southern Philippines." In: *The Australian Journal of Linguistics* 36.1, pp. 131–163.

————— (2018a). "The art of not being legible: Invented writing systems as technologies of resistance in mainland Southeast Asia." In: *Terrain* 70, pp. 1–24.

————— (2018b). "The invention, transmission and evolution of writing: Insights from the new scripts of West Africa." In: *Paths into script formation in the ancient Mediterranean*. Ed. by Silvia Ferrara and Miguel Valério. Rome: Studi Micenei ed Egeo Anatolici, pp. 189–209.

Regan, Anthony J. (2002). "Bougainville: Beyond survival." In: *Cultural Survival Quarterly* 3, pp. 20–24.

Roka, Leonard Fong (2014). "The legacy of Damien Dameng, the father of Meekamui." PNG Attitude, `https://www.pngattitude.com/2014/03/the-legacy-of-damien-dameng-the-father-of-meekamui.html`. (Visited on 26 May 2020).

Smalley, William A. and Nina Wimuttikosol (1998). "Another Hmong Messianic script and its texts." In: *Written Language & Literacy* 1.1, pp. 103–128.

Street, Brian (1984). *Literacy in theory and practice*. Cambridge: Cambridge University Press.

Tryon, Darrell (2015). "The languages of Bougainville." In: *Bougainville before the conflict*. Ed. by Anthony J. Regan and Helga M. Griffin. Canberra: ANU Press, pp. 31–46.

A "Sacred Amulet from Easter Island —1885/6—"

Analyzing Enigmatic Glyphic Characters in the Context of the *rongorongo* Script

Robert M. Schoch · Tomi S. Melka

Abstract. A newly discovered artifact known as the "Sacred Amulet from Easter Island" ("EISA") displaying a limited number of *rongorongo*-like signs is brought to the attention of *rongorongo* (RR) scholars and other interested readers. Unfortunately, the names of the original Rapanui creator of the artifact, and its first European collector, have not come down to us; however, according to an old label attached to the object, it was collected in 1885 or 1886.

Initial probability estimates of this odd-looking piece suggest a custom-built "lunar-based calendar," possibly designed for propitiatory and/or divination ends. Given the re-occurrence of the "full moon" glyph /152/ on this artifact—among other glyphs—, the best way to evaluate its function and its general meaning is by comparison with the apparent "Lunar Calendar" (C*a*6—C*a*9 [= C*r*6—C*r*9]) on tablet "Mamari" (Barthel, 1958; Guy, 1990). Although there is no exact match, the partial overlap between the "Lunar Calendar" on "Mamari" and the "Sacred Amulet from Easter Island" is of interest, and requires proper documentation.

The fact that the "amulet" may post-date the year 1864—the chronological boundary marking the arrival of Christianity to Easter Island and the generally presumed "end" of the classical RR scribal tradition—does not diminish its status as a carrier of *rongorongo* signs. "EISA," evidently, cannot be equated with the pre-missionary extended texts appearing on various skillfully carved RR objects; yet, it may be an item for possible inclusion in a special sub-corpus dealing with the post-missionary pieces.

As a corollary, we conclude that in the light of past and current RR research, hypotheses should not go by unquestioned and should be critically assessed within the available evidence.

Robert M. Schoch
Institute for the Study of the Origins of Civilization (ISOC),
College of General Studies, Boston University, Boston, MA 02215, USA
E-mail: schoch@bu.edu

Tomi S. Melka
Las Palmas de G.C., Spain
E-mail: tmelka@gmail.com

> *Omne ignotum pro magnifico est* [Everything un-
> known is taken to be magnificent].
>
> (P. Cornelius Tacitus. *ca.* 98 CE. *Agricola* (*De vita
> et moribus Iulii Agricolae*). Book 1.30)

1.   Introduction / Aims of the Study

The study of poorly known scripts that have thus far eluded generally
agreed upon decipherment can be hampered by such factors as a small
corpus for study, lack of a true bilingual text, apparent scribal varia-
tions and eccentricities within the corpus, questions as to whether the
script under consideration is an "early script"[1] or a more developed and
standardized script, and disagreements among modern scholars as to
which inscriptions should be regarded as canonical (that is, authentic)
and, thus, worthy of serious study. These considerations play into the
study of the indigenous *rongorongo* script of Easter Island (Rapa Nui),
first recorded in 1864 by the lay missionary Joseph-Eugène Eyraud.

There is a further aspect of *rongorongo* studies that should be taken
into account. The majority of texts explored in the literature occur on
wooden tablets and lack any type of specific context or supplementary
non-linguistic data. For instance, there are no known *rongorongo* inscrip-
tions accompanying indigenous illustrations, nor have many *rongorongo*
inscriptions survived on artifacts of a functional nature beyond mere
tablets; among the generally agreed upon canonical corpus of twenty-
five *rongorongo* texts (Barthel, 1958; Fischer, 1997), one is inscribed on a
long "staff," two are inscribed on *rei miro* (wooden gorget-like ornamen-
tal artifacts), and one short and partially defaced inscription occurs on
a statuette of a *tangata manu* (birdman). The remainder are on wooden
"tablets" of various shapes, sizes, and preservation status. Thus, in gen-
eral, the artifacts that record the *rongorongo* texts provide little in the way
of clues as to the meanings of the inscriptions.

An artifact related to the traditional *rongorongo* (= RR) practices on
Easter Island (Rapa Nui) was recently located in a private collection
(Schoch and Melka, 2020a). This ellipsoidal-shaped relic has an old
paper label on it (see Fig. 3 below) that reads "Sacred Amulet from
Easter Island—1885—" (abbreviated here as "EISA" [Easter Island Sacred

---

1. The designation "early script" is a convention on our part; we do not "con-
done" / endorse a teleological linear scale for the classification of scripts (cf., among
others, Moorhouse, 1946, p. 17; Gelb, 1963, pp. 190−205, who did overtly make such
claims). We recognize that the "early script/s" designation can be potentially ambigu-
ous, as it may hint at the alphabetic script/s as the epitome of perfection, which they
are not in our assessment.

Amulet]; see Fig. 4); possibly the date can be interpreted as "1886". The "EISA" object is made of painted wood with hair and two pieces of bone attached (tied) to it. What caught the early interest of one of the authors (RMS) was when he was told, before seeing it, that "All around the wooden body are various strange symbols of creatures and other geometric patterns". Direct examinations of the object - combined with high resolution photos of "EISA"—revealed a number of "scrambled" signs. Many of these resemble various *rongorongo* sampled glyphs (see Jaussen, 1893; Ross, 1940; Butinov and Knorozov, 1957; Barthel, 1958; Fischer, 1997), with the rest appearing to be geometric "decorative"-like designs, and there are also some illegible or obscured areas. Although "EISA" seems to be, at least to date, an "unicum," another old artifact from Easter Island depicted in Fig. 5 (and see also Fig. 6) can draw striking parallels in terms of the elongated / ellipsoidal shape and/or its perceived function(s): a propitiatory amulet intended to increase the fertility of sea-birds' and/or sea-turtles' eggs.[2] In any case, while this artifact is briefly and heuristically described in the legend of Fig. 5 and in footnote 2, the focus of our study is the "Sacred Amulet from Easter Island—1885/6—".

The salience and re-occurrence of glyph /152/ 🔯 on "EISA" (Fig. 1)— up to now a *hapax*[3] in the surviving corpus[4] and associated by scien-

---

2. This pre-missionary object is made of "wood" or some kind of carved plant material, and it is hollowed out so that it forms a small "container" that at one point held miscellaneous bird bones. So, in a sense it might be thought of as an artificial "egg". The symbol on side (a) appears to be a very stylized Make-make face, with side (b) portraying a strangely shaped "sea turtle"-like design (see Fig. 5). In our view of the matter, either symbol rather than rendering service to the authentic *rongorongo* script appears to fit in an iconographic context. Intuitively, however, one cannot neglect the fact that "Make-make"-like glyph /513/ 🔣 (plus, variants) and "sea-turtle"-like glyphs of class /280/ 🔣 and /290/ 🔣 are part of the *rongorongo* sign inventory (Barthel, 1958). Following this context, one may see, e.g., Geiseler (1995, pp. 65–66), "[The chief god] Make-Make is mainly represented through the sea bird eggs of Mótŭ núǐ, located on the South-West side of Rana Kao Crater; these eggs may be gathered only in the months of July, August, and September. During all other seasons they are tabu. Make-Make is worshipped through the figure of a carved or painted sea bird; examples are presented in Plates 15 and 18 [Figs. 16 and 19]".

3. Also known as V (1, N), a word (= type) with frequency 1 along the text length. On problems that very low frequency terms present for statistical and linguistic studies, see van Rijsbergen (1979); Baayen (2001); Baroni (2006). This position is briefly formulated, e.g., in McEnery and Wilson (2001, p. 77), "At the same time quantitative analysis also tends to sideline rare occurrences".

4. Cf. Barthel (1958, pp. 118, 245), "Bemerkenswert ist das Zeichen 152 für den Vollmond: in einer ovalen Umrahmung sitzt eine Figur über drei gekurvten Bögen. Anscheinend wird damit ein „Mann im Mond" dargestellt" [Worthy of perception is sign 152 standing for the Full Moon: in one oval frame, a figure sits upon three warped arches. It seems that a "Man in the Moon" is portrayed therein], and Guy (1990,

tific extrapolation or tautology with the "full moon" in the ancient Rapanui lore[5]—is more-than-enough reason to study and detail the "Sacred Amulet from Easter Island" in the *rongorongo* literature. Specifically, the only recorded instance of /152/ on C*a*7 (= C*r*7)[6] is part of the so-called "Lunar Calendar" (= "LC") on tablet "Mamari" (see Figs. 2 and 10). As both sequences ("LC" and "EISA") share # /152/, a theoretical probability may be assigned to the occurrence of each possible next glyph on the latter artifact (scattered as they are), in line with the larger setting of "Mamari". We are also amenable to the information found in the psychological literature that "word [= glyph] familiarity depends not only on frequency of perception, but also on the relevance and familiarity of a word's [= glyph's] meanings (cf. Le Ny and Cordier, 2004). Another plausible mechanism was described by Wettler, Rapp, and Sedlmeier (2005), if a person perceives a stimulus word [= a stimulus glyph], other words [= glyphs] are evoked in their memory, which are called associations".[7] This issue will receive due attention in Sections 3 and 4.

Concerning the terminology in use here and in other articles by Melka and Schoch, we call special attention to "pictorial" and "pictorial"-like in the context of the *rongorongo* script and of various writings systems in general. The first term is fittingly described by Ernst Pulgram (1976, p. 6),

> By pictorial is meant a realistic picture of something or some situation, intended to illustrate whatever message is to be conveyed. This kind of visual communication is comparable to a cartoon without caption. The translation of the picture into words is necessarily free, and does not infallibly convey the words the designer of the picture had in mind, nor do different viewers employ the same phrases or words in their attempts to render the sense.

As for the second term, while "pictorial"-*like* glyphs imply their conception by means of / like a picture / pictures, they encode (or potentially encode) a degree of linguistic information, e.g., logographic or syllabic, or mixed information, e.g., semantic and phonetic. For instance,

---

p. 136), "This glyph [→ /152/] is egg-shaped and has inside it an anthropomorphic figure sitting in profile atop a heap of rubble. All in all a very likely representation of the man or woman in the moon cooking food in the *umu* (the "heap of rubble" depicting its cooking stones), a widespread figure not only in Polynesia, but also in Melanesia…"

5. Cf. a good many authors: Krupa (1971, pp. 8, 9); Guy (1990, pp. 136–138); Macri (1996, p. 184); Fischer (1997, p. 233); Robinson (2002, p. 237); Facchetti (2002, p. 219); Berthin and Berthin (2006, p. 95); Ávila Fuentealba (2007, p. 82, Secuencia 23); Sproat (2010, p. 126); Horley (2011, p. 22, Figure 3, p. 30, Figure 9.15).

6. Numbering and nomenclature / labeling of glyphs and RR texts used herein is that of Barthel (1958). The original source for the glyphic snippets and sequences is T. S. Barthel (ibid.); however, the glyph-designs across the article are largely vectorized in line with the LaTeX format.

7. Original quotation appears in Köhler and Rapp (2007, p. 65).

FIGURE 1. A close-up image of glyph /152/ Ⓥ on the "Easter Island Sacred Amulet—1885/6—" ("EISA"). For the use of the "full moon" glyph relative to the "calendar" on Easter Island, see the clear image of "C*a*" (= "C*r*," *recto* of Tablet "Mamari") in Orliac and Orliac (2008, p. 255). Apparently, the resurgence of Ⓥ (cf. "EISA") may dismiss the status of /152/ as an isolated "exception" occurrence in the hitherto conventional *rongorongo* corpus (cf. "Mamari Tablet"). Protective gloves were used in the handling of the object; photograph © by R. M. Schoch, taken with the permission of the anonymous owner.

the Old Egyptian hieroglyphs, the cuneiform scripts of Mesopotamia, the early Chinese writing, Maya glyphs of Mesoamerica, and other ancient scripts readily dispel doubts that they were simple arrays of raw pictures (cf. Friedrich, 1971, pp. 34−51; Gaur, 1994, pp. 143−145).

It is a fine assumption that retrievable *hapax* signs may testify, among other things, "…to the author's wish to find image-bearing expressions…" (Tuldava, 2005, p. 375), which is compatible with the logographic value of the "full moon" glyph. The incidence of the pictorial-like glyph /152/ on "EISA," we point up, is (a little) too selective and specific to be a coincidence or a random artistic act. One should also keep track of the three other attested signs on "EISA"; scribal variants /V19/, /660/, and /V700/ suggest inventive shapes of their most commonly attested matrices /19/ ⫯, /670/ ⅏, and /700/ ⫰ along the corpus. This piece of evidence alone points toward the prerogative of the painter (= scribe) to reinterpret stylistically the basic designs of glyphs /19/, /670/, and /700/. Intentional (or not) different morphological realizations of a sign testify to *variants*. Such a personal hand-painting (= handwriting) on "EISA" argues for a scribal tradition aiming at standardization, yet tolerating lavish diversity due to esthetic (cf. Melka, 2014), pragmatic (the chosen medium; reduced space; interaction of paint-

FIGURE 2. (a) One of the earliest depictions of glyph /152/ is that of John Linton Palmer (1876, Plate I, 4th line). Palmer recognizes that he "...delineated a number of the [*rongorongo*] symbols inscribed on the tablets, about the same size as the originals, and taken indifferently [and, in no uncertain terms, disjointedly; *our comment*] from the casts and photographs"; (b) an "old" image of a section of C*a*7 (= C*r*7) exhibiting glyph /152/ (Thomson, 1891, Plate XLV); (c) occurrence of /152/ on line 7 of side *a* (i.e., recto) of "Mamari" (text "C") after the drawings of Bodo Spranz (in Barthel, 1958). The *hapax graphomenon* (O)motohi (= "full moon") is marked with the symbol "↓" in (b) and (c). In Jaussen (1893), subsection *Ethnographie* [Ethnographical] (d), the aforesaid glyph is rendered as "Tagata i te hare pure" [FRE. Homme en la maison de prière; ENG. Man in the prayer house] in line with Metoro's made-up reading; see, e.g., A. Métraux (1940, pp. 396–397); S. R. Fischer (1997, pp. 227–229); J. B. M. Guy (1999, p. 127, Fig. 1).

ing implement and the topology of the object), and physiological reasons (anatomical features of the authorial hand, health issues, occasional carelessness, etc.; cf. in a broader context, Schomaker and Bulacu, 2004; Davis, 2007).

Since the RR corpus in existence is limited, whether in quantitative terms regarding the corroboration of suggested hypotheses and decipherments,[8] in chronological / diachronic terms, or as to the genre variety (Melka, 2009), we are obliged to remark here: any new (pre-, or post-missionary) piece showing genuine or derivate *rongorongo* signs, merits discussion in the literature. The present focus is on the classical script, though later graphic elaborations such as *ta'u* and *mama* should be examined for their theoretical inferences, social and linguistic (see, e.g.,

---

8. The many decipherments of RR served up thus far to us represent a subject-matter that requires a special treatment elsewhere.
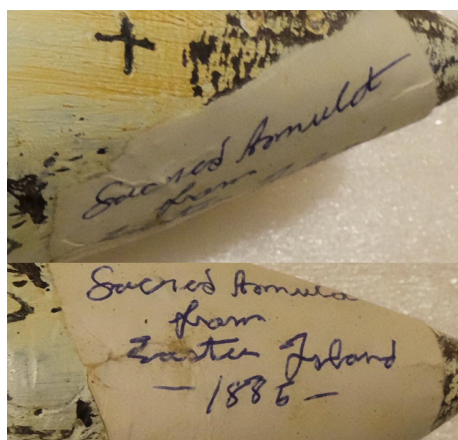
FIGURE 3. Label glued over the bottom surface of the artifact bears the inscription "Sacred Amulet from Easter Island—1885—" (or possibly, 1886). The makeshift label was made—most certainly—after the acquisition of the item. It is not known at present if the handwriting belongs to the original European collector or purchaser of the piece in question (= an "Irish missionary"), to a possible later (unknown) owner, or to Harry Geoffrey Beasley (1881–1939), another person in the line of ownerships (v. infra). At this time, any suggestion regarding additional *rongorongo*-like symbols hiding in the area beneath the paper label (conveying the "identity" and provenance of the artifact) is undecided. The owner of the artifact does not want to attempt the removal of the label. Photographs © by R. M. Schoch, taken with the permission of the anonymous owner.

Fischer, 1997, pp. 6, 513; Wieczorek and Horley, 2015; Horley, Davletshin, and Wieczorek, 2018).

The recovery of the piece discussed in this article, i.e., "EISA," is attributable to the searches and contacts of RMS with private collectors of Oceanic / Polynesian artifacts. Along with the "*Rangitoki* bark-cloth fragment" (Fig. 7) and the "*San Diego* Tablet" (Fig. 8),[9] in the absence of scholarly concern and diligent pursuit, these artifacts would probably have a nearly zero chance of coming to the notice of researchers, script experts, linguists, anthropologists, and other students. Rather, they would remain hidden in the recesses of private collections and/or antique shops.

Because of the limited number of genuine *rongorongo* items, and the dim prospects of finding other suitable and reasonably long pieces, it is in our opinion useful to peruse the "Sacred Amulet from Easter Island—1885/6—" and other objects that presumably bear genuine RR glyphs.

_____

9. Schoch and Melka (2019); Melka and Schoch (2020a).

(a) view 1   (b) view 2   (c) view 3   (d) view 4
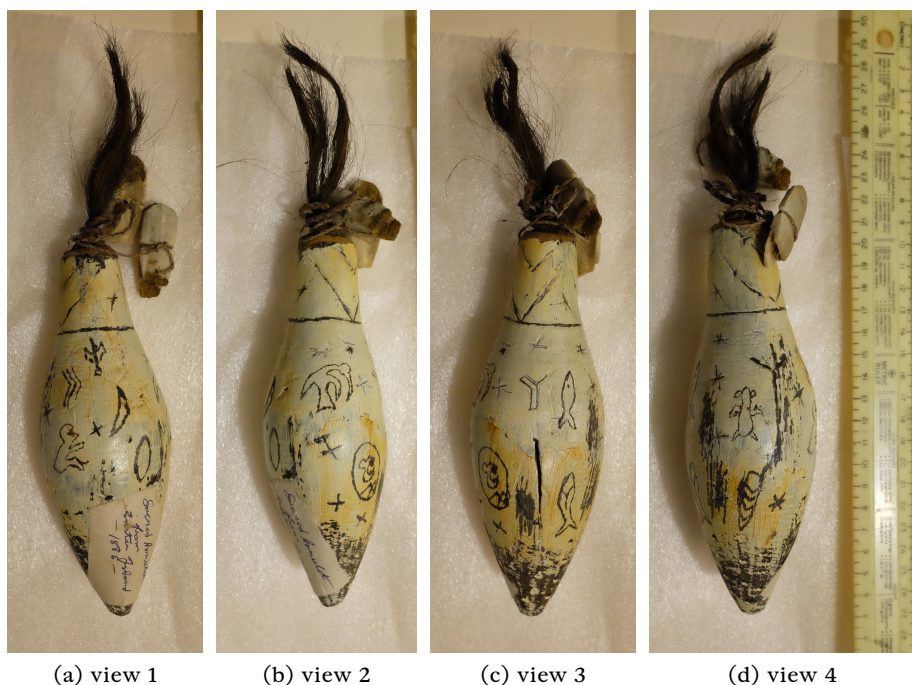
FIGURE 4. Easter Island Sacred Amulet ("EISA"). Protective gloves were used in the handling of the object. Scale is in centimeters. Photographs © by R. M. Schoch, taken with the permission of the anonymous owner.

This reemphasis is justified, if we recall at this juncture, the attitudes of some researchers on this critical matter.

Ormonde Maddock Dalton, one of the least quoted original sources across the *rongorongo* bibliography, pointed out that arguments regarding the meaning and interpretation of *rongorongo* would be stronger "…if we had really a large number of tablets to work from instead of less than twenty" (Dalton, 1904, p. 5). Sebastian Englert, in monitoring the post-1864 habit of the natives to reroute or hide RR tablets and other heirlooms in caves, explains,

> The traditions tell that the openings of such caves were carefully blocked up and concealed by their owners, and very few that could be classified as such repositories have been found. No new *ko hau rongorongo* have been discovered recently, and it seems likely that any that might turn up in the future would be in too bad condition to be of much use. (Englert, 1970, p. 78)

A few pages later, the German Capuchin priest continues,

(a)   (b)
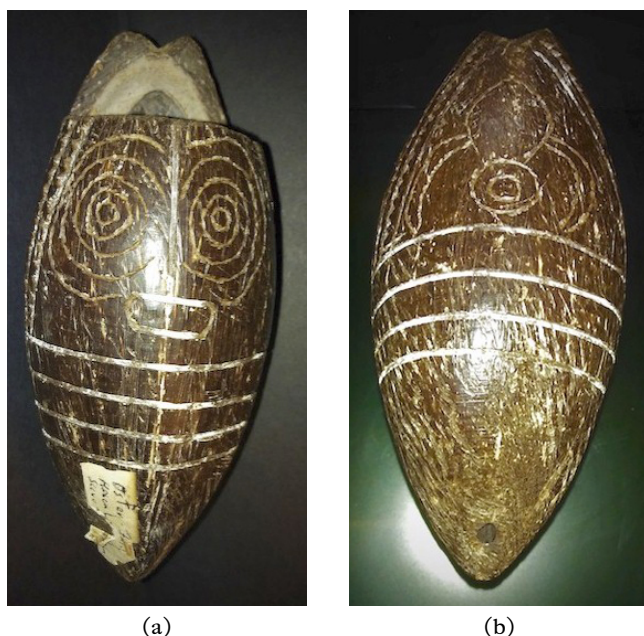
FIGURE 5. Elongated / ellipsoidal artifacts in the guise of sea-bird eggs, of *tahonga*(s) (*Tahonga* are "egg"-shaped / "coconut"-shaped / "cardiod" wooden pendants that were worn as insignia of rank / social status among the Old Rapanui residents; see, e.g., Chauvet, 1935, Plate 35, Figs. 91, 92, 93, and 94; Heyerdahl, 1975, PLATE 51a, 51b, 52b, and 52d; Orliac and Orliac, 2008, pp. 196–226; and Fig. 6) or gourd-like fruits, were apparently manufactured on Easter Island more often than some may realize. This indigenous portable object collected *circa* 1815–1816 by Otto von Kotzebüe (= Kotzebue) or a member of his crew when he visited Easter Island aboard the Russian vessel "Rurick" (cf. von Kotzebue, 1825) was subsequently passed down through his descendants and a related family until it was sold in 1990 to a private collector. Size estimated at approximately 10 cm in maximum length [currently the anonymous owner is out of contact due to the COVID-19 pandemic]. The paper label at the bottom of Fig. 5a, reads "Oster - Insel... [Easter Island] (remainder uncertain)". Photographs © by R. M. Schoch, taken with the kind permission of the anonymous owner.

    Furthermore, if some thousands of tablets had been preserved instead of the pitifully few that survive, it might have been possible to carry out the kinds of comparative studies which are impossible with the tiny available remnants. (ibid., pp. 80–81)

The Chilean ethno-musicologist Ramón Batista Campbell (1971, p. 379), in turn, was of the opinion that a few other authentic pieces may be encountered in some museums and select private collections, if

FIGURE 6. Here is illustrated a type of *tahonga* or *tahonga-like* object that is composed of wood, bark-cloth, plant fibers, fragments of feathers, and obsidian (forming the bird's beak—most of the obsidian sliver is covered with bark-cloth) and pigments; see especially Eggertsson (2011, p. 120) relative to "…the perception of [a] bird or human being hatched from an egg [= a *tahonga*-like object; *our note*]". According to the records of the current anonymous owner, who purchased it at an auction in the Netherlands in 1990, this egg-shaped artifact was collected from Easter Island in 1888 and was once owned by the Dada and Surrealist artist Max Ernst (1891–1976). Erika Vogler (1989, pp. 75–76) describes the strong interest and attachment that Max Ernst had to Rapanui human- and bird-like figurines and artistic forms; this interest was translated into a number of paintings and collages realized by him during the 1920s and 1930s (see Vogler, 1989, pp. 76–77). Scale is in centimeters. Photograph © by R. M. Schoch, taken with the kind permission of the anonymous owner.

one considered the random distribution of *rongorongo* inscriptions across the different geographic coordinates—from Honolulu in Hawai'i to St Petersburg, Russia; from Santiago de Chile to Washington, DC (USA). The Belgian researcher Jean Bianco (1976, p. 17) addressed the question in these terms,

> La rareté des tablettes pascuanes est une des principales raisons qu'évoque Barthel quant à la difficulté de pénétrer profondement la thématique de cette écriture".[10]

Thomas S. Barthel (1993), after decades-long devotion to the classical script of Rapa Nui, appears to be more restrained in his optimism regarding the acquisition of new pieces,

───────────────

10. [The dearth of Rapanui tablets is one of the main reasons evoked by Barthel regarding the difficulty in gaining deeper access to the subject-matters of this script].
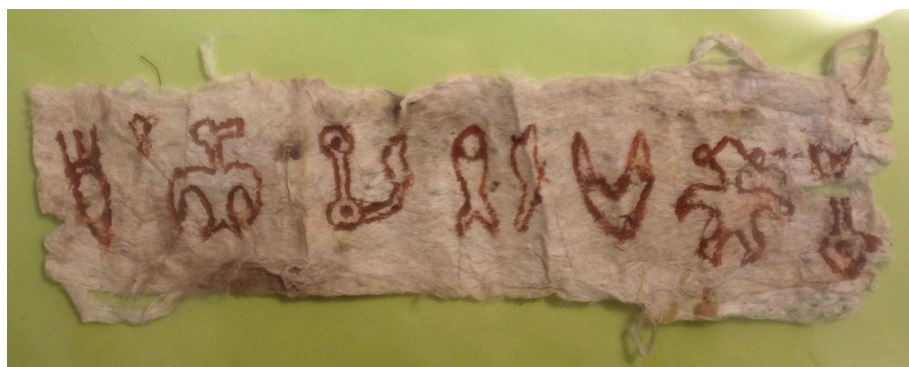
FIGURE 7. The "*Rangitoki* bark-cloth fragment". This piece was collected on Easter Island in March 1869 (Schoch and Melka, 2019; 2020b). Overall length of the fragment is approximately 15.5 cm. Photograph © by R. M. Schoch, taken with the permission of the anonymous owner.
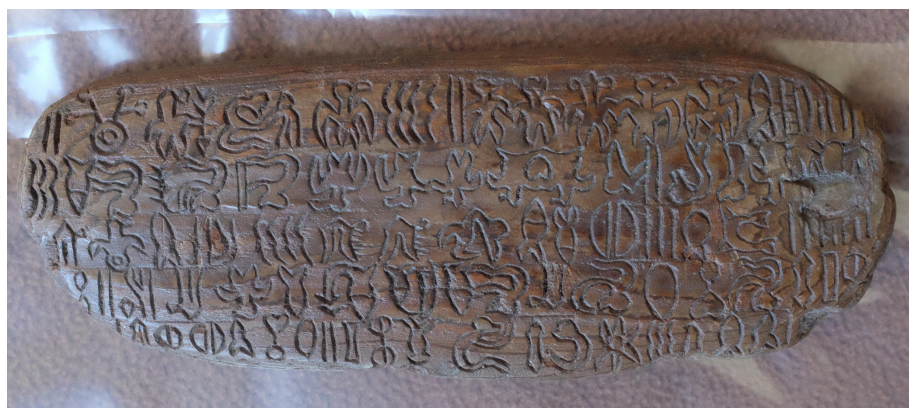
FIGURE 8. The "*San Diego* Tablet"; side *a* (?) / *recto* (?); designation of sides is out of convenience (see especially Melka and Schoch, 2020a, p. 491, fn. 14). This piece once resided in a San Diego (California) estate; it may date to the late 1850s − early 1860s or shortly thereafter (Melka and Schoch, 2020a). Overall length of the wooden tablet is approximately 16.7 cm. Photograph © by R. M. Schoch, taken with the permission of the anonymous owner.

However, certain limitations in our knowledge continue to persist. The inventory of the "Corpus Inscriptionum Paschalis Insulae" comprises only an accidental fraction of what there was in the Rapanui's dwellings at the arrival of the missionaries. Bold optimists dream of tracking down secret caves in which wooden tablets, well wrapped and protected by dry storage, await the modern discoverer. One can also nourish the secret hope that past travellers' undiscovered legacies might include *rongorongo* tablets. For the time

being let us content ourselves with the maximal evaluation of what is avail-
able. (Thomas S. Barthel, 1993, p. 174)

S. R. Fischer (initially in 1993, and later in 1997), anticipating that
Barthel's catalog (1958) of the alphabetically coded *rongorongo* artifacts
(→ A • B • C … Y • X • Z) was doomed to "failure" in case "new *rongorongo*
text[s]… were to appear,"[11] presents his own brand of catalog in which
the texts follow an orderly progression → RR 1, RR 2, RR 3… up to RR 23,
RR 24, RR 25, and highlights, "I have preferred numbers in the event—
unlikely as it may be—that more authentic *rongorongo* artefacts may be
discovered in the future" (Fischer, 1997, p. 649).

For reasons not hard to guess, the author is open to both possibili-
ties: on the one hand, to a rigidly closed RR corpus, and, on the other
(= the optimistic view), to an expandable corpus due to "good luck" new
findings. M. de Laat (2009, p. 4), another researcher caught up in the
decipherment of RR signs, conveys again the problem,

Another obstacle for the decipherment of *rongorongo*, and probably the
most severe one, is formed by the fact that the amount of the text material is
very limited. There are only 11 surviving objects which have texts consisting
of more than 300 signs. Barthel [= 1958, p. 165] estimates a grand total of
only 12,000 signs. As three of the longer texts [= the "Great Tradition" texts]
are basically the same, and other tablets share a number of passages as well,
the available material is even further reduced.

A logical deduction at this moment is that enlarging strategies concern-
ing the current corpus would have to primarily tap into private collec-
tions, still holding a great potential for new discoveries.

How large a corpus the RR script should have in order to claim or
defend any viable proposal is difficult to say. We should consider prac-
tical matters (e.g., allographic observations; falsification / revision of
a hypothesis; segmentation issues; study of glyphic transpositions; in-
spection of a particular subset related to a specific genre, etc.), though
as things stand, it should be acknowledged that size definitely matters
in the *rongorongo* studies.[12] In contrast to modern corpora that are open
to periodic updates and additions, the RR sample corpus has been (until
recently)[13] static or on the point of being "closed" (Melka, 2009). With-
out the benefit of enriching it, the RR records would most likely remain

---

11. Wieczorek (2013, p. 5).

12. Guy (2006, p. 65) in analyzing the values of a number of glyphs within the "Lu-
nar Calendar," advises, "The accumulation of hypotheses in the foregoing discussion
demonstrates how unlikely it is that *Rongorongo* script will ever be fully deciphered.
Each hypothesis has to be verified and for that a much larger corpus is needed than
what we have".

13. Here we refer to the "*Rangitoki* bark-cloth fragment" and the "*San Diego* Tablet"
mentioned earlier.

an incomplete representation of the former activity of the Rapanui people as reported by Eyraud (1866; cf. Thomson, 1891; Routledge, 1919). Present-day experts of corpus linguistics would otherwise draw attention to them as being unrepresentative and unbalanced (cf. Biber, 1993; McEnery and Wilson, 2001; Hunston, 2002), impacting, as a result, the validity of *rongorongo* studies.

At this stage, the current authors are pleased to introduce into the *rongorongo* literature the "Sacred Amulet from Easter Island—1885/6—," documenting and illustrating it.

In the future, it may be the case that other studies will concentrate on particular features, and contribute accordingly to the better understanding and interpretation of "EISA".

2.   Background

In March 2019 one of the authors (RMS) was contacted by a Canada-based antiques dealer with whom he is acquainted regarding a "collector in Hawai'i" who had a wooden "Easter Island Sacred Amulet" that was "collected in 1885" bearing "various strange symbols of creatures". Neither the dealer nor the Hawaiian collector associated the symbols on the object with *rongorongo*-like glyphs. The Hawaiian collector wanted to part with this item in order to be able to acquire some Marquesas Island artifacts in the possession of the dealer. Apparently the two men traded various items, including "EISA"; subsequently RMS was able to acquire access to "EISA" for the purpose of scholarly study (the object is currently stored in an undisclosed location).

According to the information obtained by the antiques dealer, the Hawaiian collector acquired "EISA" in London in 1985 from an English collector who stated that the artifact had previously been owned by the English anthropologist and eminent collector Harry Geoffrey Beasley (1881–1939), also known for establishing with his wife, Irene, the Cranmore Ethnographic Museum in Chislehurst in Kent (see Waterfield, 2006, pp. 78–91, and Carreau, 2010, p. 42). Reportedly the artifact was originally collected by an "Irish missionary" during a visit to Easter Island in 1885 (or in 1886, depending on how one interprets the date on the paper label, see Fig. 3). Pursuing the Beasley connection, we discover that he made "extensive acquisitions from the London headquarters of the Melanesian Mission and the London Missionary Society" (ibid., p. 44)—hence, aside from the British dealers and auction houses, it can be speculated that possibly "EISA" was acquired from one of these institutions.

Ultimately, his large collection of ethnographic material (10,000-plus objects from all over the world) "…was dispersed after Beasley's death (1939), the bulk of it being donated to six British museums between 1941

and 1955" (Carreau, 2010, p. 41).[14] In the intervening years between the Irish missionary's and H. G. Beasley's acquisition of the piece, we cannot dismiss another shift in ownership. Similar gaps are noticed after Beasley passed away: did "EISA" end up in the hands of another anonymous person (?) / or in an institution (?)[15] before reaching the English collector (= responsible for its sale in 1985)? Without hard evidence, however, readers should consider the above suggestions until proven or rejected.

As a parenthesis at this point: several consulted travelers' and ethnographic sources reveal that the post-1864 Easter Islanders were actually too happy to please or cajole outsiders by improvising chants, releasing information about the "old times," or by exchanging their *rongorongo*-inscribed artifacts for money and other useful bits and pieces—clothes, hats, hand-held weapons, glass bottles, metal utensils, and so forth. As a result, more than one scholar may come to be suspicious of items sold / traded on the antiquities market with no verifiable point of origin. In addition, a number of travelers and scholars alike have admitted to having been duped by the ingenuity of the natives in creating and selling such items.[16] In theory, we may not know with surgical precision if "EISA" was made to oblige the anonymous Irish missionary's quest for artifacts (therein, being a "knock off replica" in the last two decades of the nineteenth century), or if it came up by accident during the missionary's visit on Easter Island (→ a crude specimen of the late nineteenth century deriving from the original RR tradition). Amid the "high hopes" and "muddled confusion"—typically shrouding the documentation of many RR artifacts and their contents (pre- and/or post-missionary; authentic and/or self-styled)—very few reports can be totally verifiable. Suffice to examine the notes given in Heyerdahl (1965), Fischer (1997), Kaeppler (2003), and Hooper (2006). Under the premises, the sort of information that should really be elicited in the case of "EISA" must concern the physical object itself; the structure of the glyphs; and its alleged function considering the socio-religious setting of Rapa Nui during the late nineteenth century. These suggested lines of investigation have a better chance of yielding some results than by doing nothing in this respect, or by sitting and splitting hairs a priori.

Hence, being well worth it, we proceed here with the first concern: the description of the physical object. The main body of the object is

---

14. Cf. also Hooper (2006, p. 72), "Beasley's collection was donated after his death in 1939 to major museums in Britain—the British Museum, Liverpool, Oxford, Cambridge and Edinburgh".

15. The missing Original Beasley Collection Number (= ID no.) assigned to "EISA" would have been helpful in this context.

16. Katherine Routledge (1919, p. 271) and Alfred Métraux (1957, p. 185), in true informative fashion, offer accounts on such practices.

composed of painted wood. Attached to the top is a tuft of hair (species not determined) and two short strings tied to pieces of bone (species not determined). The wooden portion of the object is approximately 17.5 centimeters in maximum length. It was engraved and scratched so as to carry some incised symbols or decoration, especially at the top. Around the bottom of the object there are various holes or indentations. These apparent holes may be from nails or screws that were subsequently broken, cut off, or removed from the object; if so, either portions of the screws appear to remain in the object or the holes were filled with some substance. Subsequently, the wooden object was painted with a whitish to light yellow paint. Upon the background of the whitish paint the *rongorongo*-like glyphs, which are the focus of this study, were painted using a deep brown to black paint.

The wood of the "EISA" object remains undetermined. The wood is covered with paint; to be able to observe directly the characteristics of the wood would necessitate removing some of the paint and patina and cutting into the wood, which would damage the object.

Furthermore, the species of wood does not substantially affect the evaluation of the object's age nor its authenticity. Prior to initial European contact in 1722, the Rapanui collected and valued driftwood which washed up on their shores. With European contact, wood and wooden objects of various types were acquired and utilized by the Rapanui. Relative to possible sources for the hair and bone, while in pre-contact times human hair and bone were sometimes used to create objects—for instance, various fishhooks and harpoon heads were made of human bone;[17] the *rongorongo* tablet known as *Échancrée* (Barthel, 1958, pp. 18−20 → "text D" / Fischer, 1997, pp. 419−422 → "RR 3") was originally wrapped with a cord made of human hair (Orliac and Orliac, 2008, pp. 257−259)—large domestic mammals such as horses (*Equus caballus*), cows (*Bos taurus*), sheep (*Ovis aries*), and pigs (*Sus scrofa*) were brought to Easter Island after initial European contact (Ayres, Saleeby, and Levy, 2000). Distilling information from the greater Polynesian cultural family, Wallin and Martinsson-Wallin (2001, p. 8) draw a passage from Harry Geoffrey Beasley's book *Pacific Island Records: Fish Hooks*, which would seem to correspond to the ancient Rapanui's practice, too: "In many parts of Polynesia human bones had a high value, because they were supposed to contain *mana*, making them powerful materials from which to fashion tools" (Beasley, 1928, p. 50).

---

17. Descriptions and illustrations are found, e.g., in *Te Pito te Henua, or Easter Island* (W. J. Thomson, 1891, Plate LVIII, Fig. 1, Fig. 2); *The Riddle of the Pacific* (J. Macmillan Brown, 1979, pp. 188−189); *La Tierra de Hotu Matuʻa…* [The Land of Hotu Matuʻa…] (S. Englert, 1948, pp. 259−260); *Voyage vers l'Île Mystérieuse. De la Polynésie à l'Île de Pâques* (Maiani and Quer, 1996), and in *The "Fish" for the Gods* (Wallin and Martinsson-Wallin, 2001, p. 8).

(As a side note, RMS has observed that many modern collectors of "tribal artifacts" often like to claim that bone objects in their possession are made from "human bone" without an adequate basis for such an attribution.)

3.   The "Text" of "EISA," and its Implications for the Existing Corpus

Despite the modest size of the collection of *rongorongo* glyphs on its surface, "EISA" is a valuable source for study for a number of reasons, to be discussed in the present section and the next one, Discussion.

The structuring of this section into a number of specific tasks is conducive to a methodical plan and a better grasp by the readers.

The first task (1a, b) is to register the total number of signs— identifiable or not—and cluster them as *rongorongo*- or geometric-like ("ornamental"). By "*rongorongo*"-like are understood the painted glyphs that replicate original renditions / shapes comparable to those on the tablets of the agreed upon corpus (e.g., "Tahua," "Aruku Kurenga," "Mamari"), or that mimic to a greater or lesser extent those designs / shapes. The next task along the line is (2a) *preprocessing*, as known in the technical parlance. This task implies *unscrambling* the collection of scattered glyphs—in this sense "unscramble" would mean to "normalize" them as if organizing the glyphs into a more convenient and linear form,[18] similar to what the majority of English-speaking readers would perceive as a text. The choice of words is of convenience at this time, since we do not currently know if the original author (i.e., painter / scribe) chose them haphazardly, or if they represent a sample of knowledge (whether linguistically or not coded) out of a larger body, consistent with the ancient, deeply rooted Rapanui / Polynesian traditions. At a later substage (2b), the "normalization" (= linearization) facilitates as well comparing it with other portions of *rongorongo* glyphs found in the corpus. Given the retrieval of similar fragments, an explanation is expected for such "parallels" (stage 3). Here we have to exploit the previous idea that if a newly discovered "text" or "collection of RR signs" creatively repeats (briefly or at length) the sequences found in the canonical corpus, then, we have to decide on their odds of being *genuine* or *imitative* (cf. Imbelloni, 1951; Barthel, 1963; Pozdniakov, 1996; Fischer, 1997; Wieczorek, 2013; Melka, 2017; Schoch and Melka, 2019; 2020b; Melka and Schoch, 2020a,b).

(1a) *There are twelve (12) glyphs that bear slight or large similarity with RR glyphs*. Out of these, four (4) glyphs are identified in Fig. 4(a) as [# 2, → /53^2/

---

18. Terms are based on Jurafsky and Martin (2018, p. 10).

〘, a variant of Barthel's coding number /52/],[19] [# 3, → /41/ 〗], [# 4, → /380/ 〘], [# 5, → /22/ 〘]. One glyph [# 1] resembling a "flying crea-ture" / "diving wings"[20] → 〘 (Fig. 9) is not explicitly identified, though, at first glance, it purports to be painted after a glyph of the /600/ "bird"-like series, perhaps. Taking the above descriptions into account regard-ing this obscure glyph /?/—being a "flying creature" / "diving wings"—it strikes us that this glyph may represent schematically a diving frigate bird where it is diving "head-up"; that is, it is diving toward the top of the "EISA" object (see, e.g., a frigate bird in "diving stance" in Horley and Lee, 2012, p. 15, Figure 12g).[21] However, the iconographic associa-tion of this glyph with a "flying creature" or "diving wings" may also be called into question. Hence, in aiming at another solution we may en-tertain it as a variation of glyph /33/ (see a realization on the "Small San-tiago Tablet" [G*v*1] → 〘). Consider that glyphic shapes of the type /33/ appear to be related to those of the type /32/ (for a number of occur-rences of /32/ across the *rongorongo* corpus, cf. Melka and Schoch, 2020a, p. 517). Furthermore, the singular glyph-form /32/ 〘 revealed on the "*San Diego* Tablet" (ibid., Figure 14, pp. 528, and 530) shows again that *rongorongo* scribes or copyists never ran out of creative impulses or spon-taneity. In Fig. 4(b) there are two identifiable glyphs: [# 6, a "look-alike" of /660/ 〘, a variation of /670/ 〘] and [# 7, → /152/ 〘]; some discus-sion of # /660/ and/versus # /670/ will be given below. In Fig. 4(c), im-age search provides glyphs [# 8, → /44[1]?/ 〘]; matrix-glyph [# 9, → /700/ 〘]; and variant-glyph [# 10, → /700[2]/ 〘]. Further visual query reveals in Fig. 4(d) glyph [# 11, → /280/ 〘] and the scribal variant [# 12, → /19/ 〘].

(1b) As for the simple geometric-like symbols, such as the plus-shaped (+), x-shaped (x), and asterisk-like (∗) symbols, the sequence amounts in numeric terms to approximately (as some may be obscured or painted over) <2—6—5> relative to the views seen in Fig. 4 (not counting any such symbol twice, as there is overlap among the photos).

Apparently, one may think the foregoing suggests a sense of adornment—mere decorative patterns according to the whims of the original author. Nonetheless, these signs, especially those that resemble the "asterisks," could have been a *mimicry* or a *recollection* of the "starred

---

19. The superscript at /53[2]/ indicates the listing of variants following "Formentafel 1 (*Kennziffern 1–99*)" [Sign form plates 1 (Reference index numbers 1–99)] in Barthel (1958). The superscript pattern is similarly used in the other glyphs examined below.

20. We thank Gordon Berthin for his comments regarding this obscure glyph—suggesting alternatively that it is a variation of glyph /33/—and for coining the moniker "diving wings" to describe it.

21. Cf. the comment of Horley and Lee (2012, p. 17) where they assert, "Despite the head-up depiction, all the frigate birds shown in Rapa Nui rock art are actually in a diving stance, thus highlighting their predatory qualities".

disk" / "sun" glyph /8/ 🜲, part of the trigrams /8.78.711/ ⚏ along the cells B1—B7 (see Fig. 10; cf. Butinov and Knorozov, 1957, p. 10;[22] Facchetti, 2002, p. 204; Robinson, 2002, p. 237). (2a) Regarding a "normalization" and linearization of the text, there are many unknowns involved. With which glyph does one begin? Is it to be read from right to left? From top to bottom? Or in some other discursive manner? A possible preliminary "normalization" / linearization of the "text" is: /?/— /53^2/ —/41/ —/380/ —/22/ —/660 (= 670)/ —/152/ —/44?/ — /700/ —/700^2/ —/280/ —/V19/ ; another possibility (beginning with /380/ [23]—and "reading" from left to right and from bottom-to-top to top-to-bottom repeatedly) is: /380/ —/53^2/ —/?/—/41/ —/22/ —/660 (= 670)/ —/152/ —/44?/ —/700/ —/700^2/ —/V19/ — /280/ .

FIGURE 9. The "diving wings" glyph-form (most likely, shaped after a frigate bird) as it appears on "EISA". Photograph © by R. M. Schoch, taken with the permission of the anonymous owner.

---

22. The original observation of the sequence within the "lunar calendar" per Butinov and Knorozov (1957) is "On the Kohau-o-te-ranga tablet combination 1 (Table III) ["Combination 1" = /390.41-378y-670-8.78.711/; *our note*] is repeated seven times…" The Kohau-o-te-ranga tablet is another designation for tablet "Mamari," meaning, "Tablet of the Vanquished" (Fischer, 1997, p. 416). The name is reported in Routledge (1919, p. 249), relating the provenance and ownership of said artifact to the '*ariki mau* Nga'ara. The title '*ariki mau* should be roughly interpreted as "paramount / great chief"; cf. Fischer (2005, pp. 21–22).

23. Regarding the /380/ glyph (the "sitting man," mostly in conjunction with assorted glyphic affixes) → it could serve as a "delimiter" in various contexts, introducing the next chunk of text / chant… and also it may have other meanings in other contexts (the list-like texts I*a*, T*a*, G*v*, for instance).

(2b) Given the re-occurrence of the "full moon" glyph /152/—a con-firmed *hapax* in the corpus prior to its "re-discovery" on "EISA" and an integral part of the so-called "Lunar Calendar"—it is natural to start with this known-relevant section (Fig. 10). An in-depth discussion or revision of the "LC" is beyond the scope of the present article. Readers may direct themselves at leisure to Barthel (1958); Guy (1990); Facchetti (2002); Robinson (2002); Berthin and Berthin (2006); Ávila Fuentealba (2007, pp. 82–83); Sproat (2010); and Horley (2011),[24] and see for themselves the agreeing and contrasting points among the authors.

Although the positional evidence across the "Lunar Calendar" is larger and much more orderly arranged, we observe that a good number of "EISA" signs also occur within the confines of "Lunar Calendar" (C*a*6—C*a*9 [= C*r*6—C*r*9]); cf. Guy (1990). This tendency is too strange to be a coincidence or to be misled by chance resemblance. To put it simply, there seems some positive association between the "Lunar Calendar" and "EISA," marginal as it may appear initially.

Specifically, glyphs /380/, /41/, /660 (= 670)/, /152/, /44?/, /700/, /V700/, /280/, and possibly /53/,—since it has been observed to inter-change places with "feather" glyph /3/ attested on C*a* (= C*r*) (first glyph at cell 23, in the juxtaposition /3.40^1/; Fig. 10)—are under focus. Some of the "EISA" glyphs are not visibly matching those of the "LC" (appearing to be sloppily painted, as a result of the "paint-brush" used and the con-vex topology—consider the elongated gourd-like surface of the object—, or due to the idiosyncratic style of the author).

Following the progressive left-to-right order of glyphs along the "Lu-nar Calendar" (C*a*6—C*a*9 [C*r*6—C*r*9]), we attempt to correlate those with their "counterparts" found on "EISA" in order to better visualize them sequentially. Recall that Barthel's (1958) coding as any coding sys-tem associated with unknown symbols is *a judgement call* (Sproat, 2007), subject to close scrutiny by researchers (as it should be).

We obtain therein the tentative "matching" sequence (blue employed for "EISA"'s painted glyphs): /315y/ ≈ /380/ (both of the /300/-class in Barthel, 1958); "crescent"-shaped glyph /41/ = /41/; /V631b/[25] ≈ /660 (=670)/; /152/ = /152/; /78.40^1/ ≈ /44^1/ (element /78/ at cell 12, juxtaposed at the lower section of "waxing moon" glyph /40^1/,

FIGURE 10. The "Lunar Calendar" on tablet "Mamari" (C*a*6—C*a*9 [= C*r*6—C*r*9]). Original coding is of Barthel (1958, p. 51); amended coding as applied here is after J. Guy (1990, p. 136). The only exemption to Guy's (1990) amendment is code number /44/ (cell 12; cf. also Figs. 2b, 2c), left as said at first by Barthel. The calendar design is found at Robinson (2002, p. 236) and Brookman (2007); another calendaric arrangement is available at Anonymous (2005b). As readers may well notice, the trigram /2.700x.78/ is added at the beginning of Line 7 (→ C*a*7 [= C*r*7]; cf. cells 3–8). Barthel's (1958, Tafel „Mamari" Ca 1—Ca 7) tracings, Guy's (1990) illustration, and subsequent duplications of the "Lunar Calendar" fail to make it clear; in all probability, because it is on the bevelled edge of the tablet and most photographs don't show it. However, *rongorongo* scholars such as Fischer (1997, p. 414, RR 2a7) [→ ⳼], and even more notably Horley (2011, p. 22, Figure 3, C*a*6–7 II) [→ ⳼], include the trigram at the beginning of line 7 of the "Mamari" tablet. Trigram /2.700x.78/ is preceded by a "sleeping bird" glyph /V631b/ (equally missing in Barthel, 1958, Guy, 1990, and later duplicates), traced otherwise in Fischer (1997, p. 414, RR 2a6) as ⳼, and in Horley (2011, p. 22, Figure 3, C*a*6–7, II) as ⳼. Taken in whole, the newly added glyphs seem to be a variant realization of the "group separator" /V631b·8.78.711/ (discussed below). In our opinion, /V631b-2.700x.78/ is part of the *rongorongo* "Lunar Calendar" and it ought to be shown in any comprehensive review, for the sake of epigraphic thoroughness. The current partition in six horizontal blocks and the attached numeric coding is made by the present authors for evaluation purposes.

is usually appended in the upper section of the "starred disk" / "sun" glyph /8/ along the cells B1-B7 in Fig. 10). Yet, the painted glyph on "EISA" appears to be a mirror image of variant glyph /44^1/ $\mathord{\big\langle}$ per Barthel (1958). In his article, Jacques Guy (1990, p. 135; cf. Horley, 2011, p. 30, footnote 2) transcribes the glyph related to "night 11, 'Maure'" as /78/ vs. Barthel's (1958) coding /44/. Further ahead, Guy (1990, p. 144) suggests that "…glyph 78 could serve here the same function as in groups 8.78.711 and 8.78.711x…" (subscript "x" stands for an "inverted sign" in Barthel's 1958 alphanumeric code). While the original shape differs somewhat either from glyph /78/ or /44/ (see Fig. 2b, c, under "red arrow" (first arrow from the left) symbol ↓, and Horley, 2011, p. 30, footnote 2), this is unsurprising. With the *rongorongo* system being in *a state of flux* (cf. Melka, 2014), scribal variants abound here on tablet "Mamari" or elsewhere; see Anonymous (2005a); Harris and Melka (2011a); Melka (2017). The remark does not mean that trained and semi-trained scribes had no grasp of priorities when creating or copying texts, rather it points to the clear individuality and resourcefulness in the process. Better still, the crudely painted mirror image of /44^1/ $\mathord{\big\langle}$ appearing on "EISA," might hint at Barthel's original transcription as # /44/.[26]

Next, we consider the "hanging fish" glyph /711/ and /711x/ = /V700/ (a "one-eyed fish"-shaped variant painted on "EISA"); compare this with the "fish"-shaped variant /700^2/ = /700^2/. Now that mention has been made of these glyphs, the comparison would be "eased" if we consider that the C*a* (= C*r*)-"calendar" depicts two kinds of spatial orientation for the "hanging fish" glyphs: "fish up" /711/ for "waxing moon," and "fish down" /711x/ for "waning moon" (see Guy, 1990, pp. 140–141). These distinctive and cleverly conceived logograms are missing on "EISA," due perhaps to the reduced space and/or the inclination of the painter / scribe to establish tacitly their orientation and relationship with the rest of the signs. The clear similarity between the "sea turtle"-like /280/ (cf. cell C, Fig. 10) and /280/, a recognizably pictorial-like sign on "EISA,"[27] is, at this point, unassailable.

---

26. In Barthel (1963, p. 430) glyph /44/ is "read" as "kava" [the shrub *Piper methysticum*], an inexistent (or unreferenced [?]) ginger species in explorers' pre-1864 observations regarding the island (cf. Thomson, 1891, p. 464; Lehmann, 1907, p. 260; Gusinde, 1922, p. 326; Métraux, 1940, p. 159; Heyerdahl, 1965, p. 381; Fischer, 1994, p. 430). T. S. Barthel, possibly, wanted via this "reading" to trace back the origin of *rongorongo* to extra insular sources, falling within the Polynesian orbit nevertheless. Consider that the "reading" in question dates back to Metoro's chants; see especially Métraux (1940, p. 397, Figure 56, # 18).

27. For assessments of the "*ho'onu*" glyph regarding the depicted lunar cycles: see Guy, 1990, p. 145; Berthin and Berthin, 2006, p. 95, Figure 5; Horley, 2011, p. 25, Figure 5, p. 36.

As noted earlier, "EISA's" glyph /53²/ ⿰ might have a parallel in the "Lunar Calendar" in the single "garland" glyph /3/ (cf. cell 23). T. S. Melka (2013, p. 127; cf. ibid. cross-references) comments on the exchange of "delimiters" /1.53/ ⿰ and /1.52/ ⿰ as scribal variants on the "Tahua" tablet. On the other hand, "delimiter" /1.52/ has been observed to surrogate the "standard" delimiter /380.1/ ⿰, or /380.1/ plus suffix-glyph /3/ ⿰ (cf., e.g., tablet "Gr" in Barthel, 1958). Furthermore, Guy (1990, p. 144) notes that glyph /3/ occurs very frequently in glyphic ligatures (= juxtapositions / compounds) reserved for gods (= deities), chiefs, and treasured objects. We show interest in fully supporting his statement, especially in view of Routledge's report (1919, pp. 245–246) about the valued presence of *feather ornaments* during the ritual chanting ceremonies.

At present, our attention is returned to glyph /660/ as one of the most iconic figures in the "EISA" collection. This glyph stands most assuredly for the matrix-glyph /670/ (Fig. 11). Barthel (1958, p. 315) identified in their "sunken / slanted bird-head(s)" the notions "sleeping, death" (*moe*, in Rapanui language). In a follow-up study, Barthel (1963, pp. 407–408) restates that /660/ stands for "schlafenden Vogel" (manu moe) [sleeping bird]. He explains the key meaning of "sleep" (*moe*) as linked with "dream" and "death," with the element *manu* ("bird") indicating among Polynesians "phantasmal entities".[28] On page 435, Barthel (ibid.) tags # /660/ ⿰ under "schlafenden Vogel" [sleeping bird], and correlates it with "Seele?" [soul?]; while # /670/ ⿰ has the tag "moe" "Schlafen" [sleep, sleeping]. It's no secret that in the early period of the RR investigation (in the mid-1950s) T. S. Barthel relied in large measure on Metoro's chants (see Barthel, 1958, 1963, 1972; cf. Imbelloni, 1951; Bianco, 1976, pp. 18–19; Fischer, 1997, pp. 47–57, 227–229; Guy, 1999), their full veracity being dubious, at best. On top, he (Barthel, 1958) applied syncretic deductions regarding the Easter Island's ancient culture and other Polynesian-related ones[29] with the aim of achieving the long-awaited decipherment.[30] In order to cross-check these deductions con-

---

28. We would like to direct attention to Craig (2004, p. 65) for a comparison with Barthel's stated readings (ibid. 1958, 1963), "Birds are common in Polynesian myths perhaps because of their unique character of being able to fly through the heavens—something that most other living creatures cannot do. Because of this uniqueness, most birds are regarded as having a sacred nature, sacred enough to become the messengers of gods and, in many cases, incarnations of the gods themselves".

29. Cf. also Fischer (1997, p. 233).

30. J. B. M. Guy (1999, p. 129) is quite forthright in this context, "Il semble donc bien que Barthel, tout au désir de parvenir à un déchiffrement, a voulu croire que Métoro en avait la clef; et que ce besoin l'a amène à ne voir que ce qui servait ce dessein et à passer sous silence ou à présenter sous un autre jour tout ce qui le desservait. Ce désir de parvenir à un déchiffrement l'a poussé à une analyse sélective des textes de Métoro,

cerning glyphs /660/ • /670/, we might have to look into the repeated sequences /390.41-378y-41-V631b-8.78.711/ (Fig. 10). These sequences, strictu senso, are not a built-in part of the "moon calendar," rather than marking or separating the beginning and the end of the calendar and the beginning and the end of certain significant sequences such as the six and five nameless "kokore" nights (Guy, 1990, p. 138; cf. Horley, 2011, p. 22, Figure 3). Of special interest are here glyphs /V631b/ 🐚, /V671/ 🐚, and /V631by/ 🐚, all of them appearing as variants of matrix /670/. What is curious here is that the /670/-variants occur within "group separators" that deal with moon phases *in a state of change*, growing and diminishing, as recorded by the ancient Rapanui astronomers.[31] In this context, the "sleeping bird"-like glyph could have been suggestive of these phases, especially if the "diminishing" or "devoured" moon was perceived as entering in *a state of lethargy*, hence, *death*. We should also consider that *moon* is often associated cross-culturally with *sleep*, and its related states, *dreams* and *unconsciousness*. As it happens, the ancient Rapanui scribes experienced the same archetype and coordinated metaphorically through "sleep-like" symbols part of the moon phases.

As *rongorongo*-related studies are not straight-line journeys with convenient results for all the involved parties, we should attempt to bring at hand different glyphic sequences to support the hypothesis /660/ = /670/ (see Figs. 11 and 12). Yet, in view of space constraints, various sequences might have to be located through Barthel's (1958) conventions, along with other useful scholarly references.

For the configuration of glyph /670/ and its scribal variants, we may have to quote a formula-like sequence that appears in several environments (see especially K. Pozdniakov, 1996, p. 295, Fig. 3; R. W. Sproat, 2003; F. Ávila Fuentealba, 2007, p. 44, Figura 40; T. S. Melka, 2007; P. Horley, 2007, p. 28, Figure 3, iv; Horley, 2010, p. 53, Figure 9), plus the "Mamari" tablet, C*b*2 (= C*v*2), cf. Fig. 12. The recoding of glyph /630/—on C*b*2 (cf. Barthel, 1958, p. 52)—into /631/ 🐚 is performed after CEIPP (Anonymous, 2005a). This goes to show again that bias in Barthel's notation can be eluded through the independent structural-visual comparison of sequences. Other details such as those on A*b*4 • C*b*2 (= C*v*2) • C*b*4 (= C*v*4) • H*r*4 • P*r*4 • Q*r*4, provide compelling sup-

---

sans relever les contradictions" [So it does seem that Barthel, driven intensely by the desire to reach a decipherment, wanted to believe that Metoro had the key; and that the need led him to see only what served his purpose and to hush or to show under a different light all that contradicted him. This desire to reach a decipherment drove him (= Barthel) to a selective analysis of the Metoro's texts, without revealing the disagreements].

31. See also W. Liller (1993, p. 36) for astronomical petroglyphs possibly addressing moon calendars (in particular, the so-called *Papa Mahina* "Moon Rock" near Ahu Ra'ai, Easter Island).
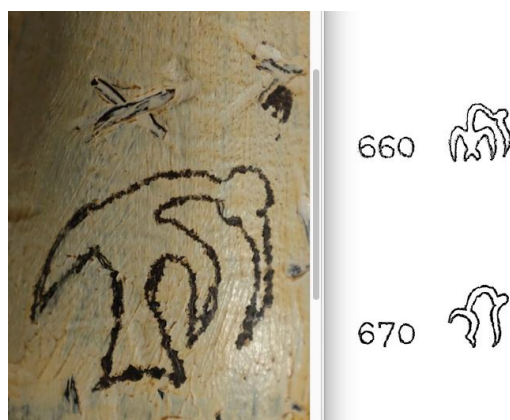
FIGURE 11.   The "bird"-like painted sign on "EISA" as contrasted with
glyphs /660/ and /670/ in Barthel's coding (Sign-form plate 7 - Reference in-
dex numbers 600–699 in *Grundlagen…*, 1958). While it is evident that the glyph
preserves its relationship to the natural referent (i.e., a flying life-form), the con-
veyable meaning is not strictly restrained univocally; various instances of early
pictorial-like scripts or modern systems endorse such a statement (cf. Houston,
2004a; Sproat, 2013, pp. 14–17). © Photograph of the leftward section by R. M.
Schoch, with the permission of the anonymous owner.

port for the variability of glyph /670/ along a fixed formulaic set (see
Cb2 [= Cv2] in Fig. 12). The formula-like string attains its most ex-
plicit form on Cb4 (= Cv4) (see Harris and Melka, 2011a, p. 140), where
the "sleeping bird" glyph features the code number /670/ 🐦. Otherwise,
on Cb2 (= Cv2) • Ab4,[32] the aforesaid "bird"-like glyph obtains num-
bers /660/ 🐦 and /680/ 🐦. Next, the Hr4 • Pr4 • Qr4 strings are posi-
tively building upon the Cb4 (= Cv4) formula set, though in a fairly con-
densed manner (Fig. 12). The "sleeping bird" glyph is not there, and the
ancient chanter had to count on the "fish"- and other "bird"-like glyphs
(with or without "excrescences") in order to retrieve their meaning. RR
corpus has many examples where the material reduction within the se-
quences (= ellipsis) does not hide from view their association with the
more complete ones.

    If researchers from a different era and sundry geographic locations
can advocate (to some extent) to their similarity, for the former *rong-
orongo* masters, we trust, it may have been a matter of simple routine.

    For a more diverse approach, we may explore the /670/ glyph-shapes
in the following nearly-parallel sequences (Fig. 13). Without wrench-
ing a translation out of them and (without) claiming traces of "lunar

---

32. For further discussion on Ab4, refer to Melka (2016, pp. 229–230, Figure 6).

"Mamari" tablet, C*b*2 (= C*v*2)

380.1- 67.10f- 67- 730- 670- 1-631- 571-700- 380.1-

"Tahua" tablet, A*b*4

1.52-   680-   65-600.65-600.65

"Mamari" tablet, C*b*2 (= C*v*2)

380.1- 600.64- 400-64-660y-64h-660-

"Mamari" tablet, C*b*4

380.1- 605.700x-605.700x-605.700x-670
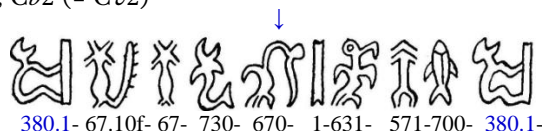
"Great Santiago" tablet, H*r*4

600:701- 600 -600.711

"Great St. Petersburg" tablet, P*r*4

600:701- 600 -400-710
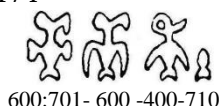
"Small St. Petersburg" tablet, Q*r*4

600:701-600s-600-710

FIGURE 12. Glyphic strings from the *rongorongo* corpus deal with "bird"-like glyph /670/ [for the more RR-demanding readers, it is recommended the careful perusal of the "Great Tradition" tablets (cf. Barthel, 1958), plus "Tahua" and "Small Washington," in search of allomorphic instances of glyph /670/ (cf. Ross, 1940; Kudrjavtsev, 1949; Guy, 1985; 2006; Pozdniakov, 1996)], the variability of the glyphic material, and the individuality of each pre-missionary hand. In the case of strings H*r*4 • P*r*4 • Q*r*4, the past scribal experience was called on to restore the missing glyphs, such as they appear on C*b*4 (= C*v*4), C*b*2 (= C*v*2), and A*b*4. The symbol "↓" points at # /670/ and its professed variants. Coding, marked in blue, stands for "delimiters," used for parsing the flow of glyphs according to specific chunks of texts. [For readers perusing the black-and-white version of this article, the "delimiters" marked in blue are /1.52/ and /380.1/].

calendars," it is interesting, however, to notice the presence of the col-
location /40-670/ ("crescent"-like—"sleeping bird"-like glyphs) along
tablets E*v*6 ● S*a*4 ● G*r*1● K*r*1−2.[33] We should not fail to remember that
RR-like signs /41 (≈ 40)/ and /660 (= 670)/ are also found on the painted
surface of the "Sacred Amulet from Easter Island" artifact.

"Keiti" tablet, E*v*6

515- 40- 670- 2.108a?-22.62-
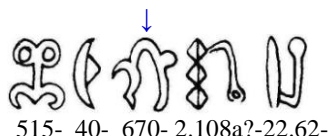
"Great Washington" tablet, S*a*4

510-39- 40- 670- 2.63- 62.1.62-

"Small Santiago" tablet, G*r*1
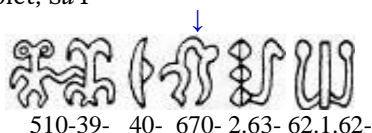
20.3- 515- 40-670- 20-22-62.1-
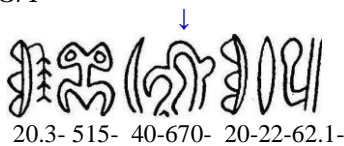
"London" tablet, K*r*1−2

20-   39- 40- 670?- 2-22- 62.1-
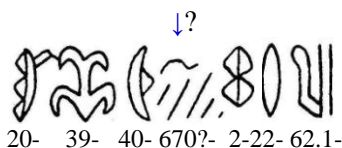
FIGURE 13. Comparison of four *rongorongo* strings (cf. Ávila Fuentealba, 2007,
p. 46).   Fused glyphs /515/ on E*v*6 and G*r*1 are in effect the individual
glyphs /510-39/, as visibly shown on S*a*4. Inclusion of # /670?/ on K*r*1−2 is
done by the present authors, in a hypothetical analogy with the rest of the se-
quences. The last compounded glyph—rendered in various forms in each se-
quence, /22.62/; /62.1.62/; /62.1/—apparently denotes a "textual delimiter".

(3) *Possible explanations*: The first question that strikes our mind is: how
often it might be expected that a hapax, i.e., glyph /152/ ⓖ, may re-

---

33. M. Harris in Harris and Melka (2011b, p. 255) presented in Table 7 a term-
to-term similarity analysis using LSA (= Latent Semantic Analysis) applied to the
whole corpus. Findings revealed that glyph /670/ favors among the first few ranks:
glyph /2/ and the "moon" glyph /40/ in line with previous observations of these
glyphs on the "Mamari" and "Keiti" tablets.

appear ca. 150 years later by chance on another artifact since the acquisition of the original "Mamari"-tablet (end of 1869 or beginning of 1870, see Fischer, 1997, p. 417)? The "re-appearance" of /152/ should be of course envisioned in its context of regular associations with the other glyphs arguing for a Polynesian-type "lunar calendar". On the flip side, we have to mention the occurrence of glyph /152/ ⓖ in the company of the compounded form /690.41/ 🜂 on an obscure "tablet," reproduced by Francis Mazière (1968, p. 64, facing). The writer (1968, p. 64, facing) provides a caption for the poorly illuminated image which reads "A toro-miro tablet covered with signs or ideographs. The writing, which remains a disturbing archaeological problem, is read by holding the tablet horizontally and turning it after each line". The examination of what is traceable from the "...signs or ideographs" reveals that this "inscription" bears entire sequences and loose glyphs from tablets "Mamari" • "Keiti" • "Tahua" • "Aruku Kurenga" • "Small Santiago," and perhaps from other ones. The fact is the *rongorongo* genres / sub-genres appearing in the authentic inscriptions (e.g., astronomical calculations, list-like formulations, etc.) are so badly and nonsensically mixed up in Mazière's "tablet" that they defeat their purpose. Specifically, glyph /152/ ⓖ which occurs strictly within "lunar calendars" so far (cf. "Mamari" and "EISA"), besides the compound /690.41/ 🜂, is combined indiscriminately with many other borrowed non-"lunar calendaric" glyphs. The observations can only lead to the conclusion: this mishmash tablet is the work of a forger.[34] The other issue here is *if* we have an attempt at trickery for lucrative purposes or an honest imitation for artistic ones. We are not picking sides at this time. Yet, for those people who are not careful while assessing the authenticity of many "*rongorongo*" artifacts we may cite the serious allegations of Grant McCall (2010, p. 49) against Mazière, "Amongst many items he acquired during his sojourn, were some skeletons from caves and other burial places, as well as commissioning a number of carvings that he sold as genuine ancient pieces".

And to return to the question on the frequency of re-appearance of glyph /152/ ⓖ, in any newly discovered *rongorongo*-related context: since the corpus is randomly collected and there is no accord on the fixed number of glyphs, we may not obtain exact figures as to the statistical calculations. A first estimation, however, tends to be close to *very, very seldom*. The curiosity of a present-day researcher is further piqued by the

---

34. A. van Hoorebeeck (1979, p. 268), while admitting some perplexity on his part, refers to the Belgian researcher Jean Bianco, who assured him that various "sign groups" on this tablet are "extracted from well-known [= authentic] tablets, especially from 'Tahua,' 'Mamari,' and 'Aruku Kurenga'". We would contend that Bianco's assessment sends up equally a red flag as to the authenticity of the "Mazière tablet". For our part, we have stated previously (Schoch and Melka, 2020b) that, in our assessment, this tablet is dubious.

fact that this *hapax disgraphomenon* V(2, N) at present, i.e., /152/, is accompanied by a cluster of glyphs that finds more than a partial "match" in the "Lunar Calendar". These two simple facts suggest that "EISA"'s mixed glyphs are neither arbitrarily nor awkwardly chosen and painted, contra the combinations on Mazière's tablet (1968, p. 64, facing). Quite the opposite, they imply a conscious selection process among the hundreds of available RR signs that the author went through, whether in pre- or post-missionary times. It is presumable, even plausible, that the author knew how to connect mentally the apparently unordered glyphs, intuit the specific function of each glyph, their interaction along the "text," plus the expected effects.[35]

Now we move on to a series of authors who comment on the practices of Old Rapanui regarding the observation(s) of time and heavens. Carl E. Meinicke (1876, cited in William Churchill, 1912, p. 336), has a pithy formulation on such activities, "Sie kennen eine Art Chronologie und bestimmen die Monate nach dem Mondsumlauf [They (= Rapanui) know of a type of chronology and order the months according to moon cycles]". Wilhelm Geiseler (1995, p. 58), in command of the warship *Hyäne*, says in this respect, "In Rapanui the designation of time is according to the various seasons of fruit ripening, i.e., their harvesting and their consumption, etc.". Carlos Charlin Ojeda (1947, p. 80) in describing a great number of caves found on Easter Island, distinguishes one called by the name "Ana Ui-hetúu,"[36] "caverna desde donde se miraban las estrellas [cavern where the stars were seen]". This site located on the western coast of the island, near Ahu Okahu, would be known in scientific terms as the "cavern of the astronomical observatory". The most obvious use of the place, according to Ojeda (ibid., p. 80), had to do with prophylactic magic,

La obligación más importante de estos astrónomos-sacerdotes era prevenir a los isleños de las influencias, benéficas o nefastas, de los astros y planetas. Así, 'Matamea' (Marte), o la constelación 'Tautoru' (Orión),[37] o una estrella 'Pau', que aparece entre Octubre y Noviembre, cuando coincidían, provocaban una enorme alarma por el significado desastroso de este acontec-

---

35. As it happens very often with particular epigraphic and archaeological samples and in broader relatable contexts, the original scribes / creators / artisans did not arrange or produce the artifacts (inscribed or not) for the convenience of future researchers / scientists. Although we may have a general idea on the meaning of an inscribed artifact, the exact nuances, the full symbolism, and extra meanings (e.g., esoteric or sexual) will remain hardly recognizable.

36. The information of Carlos Charlin Ojeda (1947, p. 80) is mostly based on the earlier report of Alfred Métraux (1940, pp. 52–53).

37. Englert (1948, p. 506) includes also the entry "*Tui* Orión (constelación astral)" [*Tui* Orion (stellar constellation)].

imiento: muertes, destrucción de los productos agrícolas, y malos frutos en la pesca".[38]

Similarly, William Liller (1993, p. 5), in noting the importance of the movements of celestial bodies, pauses and tells us that ancient islanders planted their crops according to the phase of the moon (as some of us still do today) and feared for their well-being when Mars grew bright. Next, Giulio Magli also comments on the keen observation of the heavens and celestial bodies by the ancient Rapanui (Magli, 2009, pp. 249–251).

This report would be even more sensible if one considers that pre-contact Rapanui descended from great navigators who charted their ocean routes following the stars (cf. Buck, 1938; Englert, 1948; Äkerblom, 1968; Lewis, 1972). As they settled on their new home, they were faced with a sub-tropical climate characterized by seasonal changes which had to be closely monitored. Their successful subsistence may have depended on the seasonal patterns of planting as well as on those linked with the arrival of various species of fishes, birds, and turtles (Liller, 1993). A monthly calendar that recorded the orbit of the moon by adjusting days and nights—the moon's cycle being in misalignment with the night-day cycle as well as the yearly cycle—was a requisite device to keep it aligned with the local seasons; such adeptness is shown on the "Lunar Calendar" of tablet "Mamari" (see Sproat, 2010, p. 126; cf. Englert, 1948, pp. 311–312).

By analogy, the "organized glyphic chaos" on "EISA" could have been a miniaturized version of the well-structured information on Mamari's calendar.[39] Through magical operations, it was intended to grant "favors," "control," or some "divinatory scope" to the owner over phenomena and staple commodities that concerned directly his / her (?) life: cropping and harvesting seasons; bird and fish migrations, and so forth. Although on a lesser—and a much more private—scale, it would seem that "EISA" performed along similar lines as "Mamari"'s Lunar Calendar. This hypothesis will receive further attention in Section 4, Discussion.

---

38. [The most important duty of these priest-astronomers was to intervene for the islanders relative to the influences, beneficial or ill-omened, of the stars and planets. Thus, 'Matamea' (= Mars), or the constellation 'Tautoru' (= Belt of Orion), or a star 'Pau', that appears between October and November, when in conjunction, caused a heightened alarm due to the disastrous significance of this event: deaths, destruction of agricultural crops, and poor catch during fishing].

39. Or based on some similarly inscribed "calendar" elsewhere, lost or unknown to present-day scholarship.

4. Discussion

Attempts to determine the nature of the glyphs on "EISA" warrants an
extension to Section 3. Since there is no a priori knowledge about the
existence of the artifact or its painted brief message, different working
lines are in order: the most plausible one will endure, with the ques-
tionable ones discarded in due course. A number of possibly (or not)
script-like properties related to "EISA" will assist us in achieving more
informed estimations (see detailed discussion in Sproat, 2013; 2014; and
Melka, 2017). At the outset, we treat the general layout of the glyphs,
as it may yield relevant information to the meaning of the artifact itself.
The layout (= spatial orientation) of the "EISA" glyphs clearly cancels
out the feature of linearity as observed in the *rongorongo* tablets,[40] or in
other real-world written documents. Linearity, i.e., a meaningful con-
catenation of signs / letters / characters across texts, is a major aspect
that describes writing systems (cf. Friedrich, 1971; Harris, 1986; DeFran-
cis, 1989; Gaur, 1994; Daniels and Bright, 1996; Sproat, 2000; Houston,
2004a; Rogers, 2005, p. 9).[41] We also know linearity does not compulso-
rily stand for a one-dimensional arrangement of signs from left-to-right
as in the Roman-based English writing system. Many speech-related
inscriptions show contiguous arrangements from right-to-left; top-to-
bottom; bottom-to-top; in a circle; or even more exotic forms, such as
spiral sequencing of signs.[42] However, given the distributional frame of
rongorongo-like signs on "EISA" with their proven absence of a contigu-
ous alignment, these signs can barely be viewed as an exotic literary
experimentation.[43] In point of fact, what we may have here is a "bag-
of-glyphs"—comparable to a bag-of-words, that is, an unordered set of
words with their position ignored (cf. Jurafsky and Martin, 2018). An

---

40. An object with a non-linear arrangement of *rongorongo* glyphs is the "*New York*
Birdman (*tangata manu*)," a wooden statuette that has a number of scattered glyphs
(*ca.* 35–40) grouped in seven discrete blocks (cf. Métraux, 1940, p. 256; Barthel, 1958,
p. 33; Fischer, 1997, pp. 506–508; Kjellgren, 2001, p. 46, Plate 4 Birdman Figure [*tan-
gata manu*]; Lelièvre et al., 2010, p. 135). Despite the absence of linearity, the artifact
is indexed to date along with the other artifacts in the canonical corpus of RR inscrip-
tions. Researchers generally acknowledge that the "Birdman" glyphs are difficult to
read. They were just traced out in preliminary fashion but never deep-etched via a
"shark-tooth". A revision may be due in this case, in order to glean its current status.

41. "All writing has an underlying linear organization: that is, symbols follow each
other in some sort of predictable order" (Rogers, 2005, p. 9).

42. Cf. Gelb and Whiting, 1975, p. 101; Godart and Olivier, 1982, pp. 152–153; Gaur,
1987, Gaur, 1994, p. 166; Damerow, 1996, pp. 217–218; Fischer, 1997, p. 351; Sproat,
2000, pp. 56–60; Krämer, 2003; Houston, 2004a,b; Jannot, 2005, pp. 36–37; Rogers,
2005, pp. 9–10; Massarelli, 2014; and Fig. 14 herein.

43. The works of Gaur (1994); Bantock (2000); Albright (2000); Harris (2001) in-
clude a number of such experimentations.

attempt to unscramble and reference them to a practical setting derived from the relationships of the rotation periods of heavenly bodies ("Lunar Calendar" on C*a*6—C*a*9 [= C*r*6—C*r*9]) was previously described in Section 3.

FIGURE 14. This Egyptian hieroglyphic arrangement follows a top-to-bottom pattern. From a passageway leading to the burial chamber in the tomb of Ramesses III (Userma'atre'meryamun) (1194–1163 BCE), second ruler of the Twentieth Dynasty (1196–1070 BCE), Valley of the Kings, west bank of the Nile, across from Thebes (Luxor) on the east bank of the Nile. [Note that the exact dates of the reigns and dynasties are disputed.] © Photograph taken by R. M. Schoch during the summer of 2019.

We are very much aware that even the narrative picture stories (for children or adults) are based on a sequential progression abiding by a regulated order. Although pictures per se are freed from speech, we have to follow their directionality so as to make an immediate or a reasonable mental translation.

Of further note is the unexploited painted surface on "EISA". Research has shown that RR scribes "…took advantage of every centimeter of free space" during the writing process (Harris and Melka, 2011a, p. 126, Fig. 2), at least where the indisputably pre-missionary extant

tablets are concerned (Métraux, 1957, p. 204;[44] Michelot and Michelot, 1979, p. 58;[45] Fischer in Dederen and Fischer, 1993, p. 182;[46] Melka in Harris and Melka, 2011a, p. 125[47]). The fact that this characteristic feature is absent on "EISA" may have to do with the unusual ellipsoidal shape of the object, or the painter's choice to map glyphs out in his own terms after the "Lunar Calendar". In consequence, the "EISA" glyphs may appear quite random-looking to a twenty-first century viewer; yet, we assume that their true spatial relationships were recognizable to the original painter / owner of the artifact.

Another observation is that "EISA," unlike the canonical RR tablets ("Mamari" included), does not consistently show repetitive signs or sign-groups. The presence of the "fish"-shaped glyphs is the only real exception, if the "geometric"-like signs are confidently ruled out as *decorative*.[48] This raises again the question of whether we should consider the "EISA" text based on linguistic patterns (or not). This brings us to another observed trait: the much reduced number of featured signs on "EISA".

The brevity of "text" is another stumbling-block related to the said lack of repetition.[49] Otherwise, for a prudent judgment that a symbol system is writing, the number of signs is important (cf. the "infamous" corpus problem in RR, as noted regularly by past and present scholars). If the "EISA" signs were studied in isolation, i.e., the reference frame of the "Lunar Calendar" or any other genuine chunk of *rongorongo* text was unavailable, it would have been considerably less easy to argue about their *meaning* and *structural relationships*. Twelve (12) identifiable *rongorongo* glyphs do not offer much leeway for a controlled interpretation, let alone a decipherment. At which point we should recollect again the markedly spatial distribution of the "EISA" glyphs. In principle, this pattern hints

---

44. "Whatever the shape and size of these pieces of wood, they are invariably covered with signs on both sides, without the slightest space being wasted".

45. "…les graveurs voulaient utiliser au maximum toute la surface disponible de cette matière si rare" […the etchers (= scribes) wanted to use to the greatest extent all the available surface of such a rare material].

46. "…each scribe apparently strove to exploit the greatest possible amount of the precious wood".

47. "…*rongorongo* artefacts were carved far and wide, with scribes filling purposefully every available spot with signs (Figure 2)…".

48. However, it may not have been the case that they were merely decorative. For all we know, the exact placement of these "geometric"-like signs may have been very meaningful to the creator and to the user of the "EISA" artifact, perhaps representing "stars" or "planets" in the night sky.

49. Similar concerns were raised earlier by Sproat (2014, p. 469), advocating that statistical measures show a negative correlation between the repetition measure of linguistic / non-linguistic units and the mean text length. Or to put it in layman's terms, "…the shorter the text, the less chance there is for repetition".

at other symbolic / decorative arrangements that do not express linguistic information, strictly speaking (tessellated pavements, woodcarving, tiling, carpet patterns, and geometric ornamental designs, cf. Jones, 1856; the Vinča "religious" symbols, cf. Winn, 1981; tilings and tessellations, cf. Grünbaum and Shepherd, 1987; the rock carving surfaces at Nämforsen, Sweden, cf. Tilley, 1991; body painting among the Xavante people, Brasil, cf. Polo Müller, 1992; rock art from Cave of the Hands—province of Santa Cruz, Argentina, cf. Gradín, Aschero, and Aguerre, 1976; Wang et al., 2010).

The hitherto script-like properties of "EISA" are not especially favorable to the "linguistic hypothesis" of its contents. Another scholar (not affiliated with the present authors) might even say that whoever painted the glyphs did not mean to convey speech at all, i.e., phonological information.

Our main objective is not to strictly speculate about the encoded degree of speech (substantial, marginal, or zero), but to seek clues about the function that "EISA" might have served in the past. It is evident that symbols / paintings (devoid of speech), or characters / syllables / words (tied to speech) do not occur isolated from a socio-cultural context because if so, then they would have no meaning. Logically, any particular symbolic scene or a piece of text is the outcome of one or more specific artists / writers, in a specific individual style / dialect of a specific language, at a specific time, in a specific place, for a specific function, to rephrase thoughts of Tilley (1991) and Jurafsky and Martin (2018).

4.1.   Possible Function of the "EISA" Artifact

While the painted glyphic content of "EISA" suggests (cum grano salis) a compacted and personal version of the "LC" ($Ca6$—$Ca9$ [= $Cr6$—$Cr9$]), we should explore if the artifact itself (shape, applied paint, elements such as the tied hair tuft and bone pieces) provides a context in some implicit or explicit way for further elucidation. "Context" is understood in this sense, as someone's construction, the conceptual environment of a text, the situation in which it plays a role (see Krippendorff, 2004, p. 33). As a parenthesis, it is also worth quoting John Chadwick (2000, p. 26) regarding methodological issues in deciphering, "A cool judgement is also needed to discriminate between what a text is likely or unlikely to *contain*". Furthermore, we admit that distinguishing between the properties inherent in the artifact and those that are part of the act of interpretation,[50] is by no means easy (cf. Elkins, 1996).

Despite having a wooden body, the configuration of "EISA" is reminiscent of a gourd-shaped or bottle-shaped object. Whether this was

---

50. In this respect, by RMS and TSM.

merely accidental or planned in advance by the original artist / owner, it remains to be seen. Gourds (*hue* in Rapanui language)—pertaining to the species *Lagenaria siceraria*[51] - have been traditionally planted and used together with other basic cultigens on Old Rapa Nui. Legend has it that gourds were brought to the island by the celebrated navigator and chief Hotu Matu'a (see the chapter *The Voyage of Hotu Matua*, in Barthel, 1978). Although one has trouble adjusting to the ancient Rapanui legends / chants and obtaining clear facts, also due to reinterpretations and possible linguistic contamination over the years (cf. "Rapanui Manuscript E," ibid.), they are still culturally worthy of studying. A case in point is the account of *ipu ŋutu* (Schnabelkürbis = beaked calabash) and *hue* (Flaschenkürbis = bottle gourd), with the term *ipu ŋutu* used instead of the real *hue*. Barthel is inclined to explain this by virtue of the name which describes the function of bottle gourds (as indispensable receptacles in the Old island culture).[52] A striking parallel comes from Thomson (1891, p. 535): in his report[53] about the gourd-vessels he collected (one, now lost, which presumably carried *rongorongo* inscriptions), he talks of the "calabash" called *Tata* and used chiefly in boats for bailing. Said object, according to the National Museum of Natural History Smithsonian's database (2014), corresponds to the access number "E129758-0 Gourd" and is indeed a "beaked calabash," being, however, of the *Lagenaria* sp. As it turns out, Thomson (ibid., p. 535) describes the second listed item called *Epu Moa*: "Known as the fowl gourd, and a superstition ascribes a beneficial influence over the chicken fed and watered from it".[54] In the NMNH Smithsonian's database (2014) it has the access code "E-129757 Gourd". What is of certain bearing for our investigation is the fact that not unlike the "engraved skulls" (or other material supports),[55] even particular gourds seem to have had propitiatory effects on the wellbeing and multiplication of chickens. There are several accounts that bear out the effect and use of magic in ancient

---

51. In Métraux (1940, p. 157) and Barthel (1978, p. 133) one finds also the Latin denomination *Lagenaria vulgaris*.

52. Barthel (ibid., p. 133). Englert (1948, p. 456) offers also the entry "kaha" in his "*Diccionario Rapanui−Español*" "calabaza, calabacina (que se usaba como vasija de agua) [calabash, vessel (formerly used as a water jar)]".

53. Similarly, in the report of George H. Cooke (1899, p. 722) concerning the "NAMES OF SOME OF THE RAPA NUI PLANTS" he gives for "hue" the translation "gourd-vine". George H. Cooke, it must be remembered, was the ship's surgeon on-board the U.S.S. *Mohican* vessel which visited Easter Island in 1886.

54. See also Métraux (1940, p. 157), "Gourds were of two kinds. The round ones were used as containers for small things and the elongated ones were for water".

55. On the supernatural power known as *mana*, superstitions and outright taboos reigning among the Old Rapanui, see Thomson (1891); Lehmann (1907); Routledge (1919); Métraux (1940); Englert (1948); Barthel (1958); Fischer (1997); Mordo (2002).

times. Whether *kai kai* chants[56] or *rongorongo* glyphs, they were endowed with *mana* which could cause helpful or destructive results in accordance with the wish of the supplicant / imprecator / chanter. Alfred Métraux (1957, p. 186) mentions with regard to string figures and their correlated chants,

> These chants dealt with all the circumstances of life, love and death. A great number of them were spells that had the power to save people in danger and to multiply plants and animals. Others were panegyrics addressed to chiefs on solemn occasions.

Thor Heyerdahl (1965, p. 366) reported that "Most of the Easter Islanders still believed that their ancestors performed supernatural activities through contact with 'devils', and that the *rongo rongo* was provided with inherent *mana*". In the same spirit, other by-products related to the accounts and speculations on "gourds" are: the "calabash… covered with hieroglyphics similar to those found on the incised tablets" (cf. Thomson, 1891, p. 535; Melka, 2017), and Barthel's (1978, p. 133) logographic assignments, "In the *Rongorongo* script there is one grapheme for calabash" (*ipu*, Rongorongo sign 74 〰) and another one for bottle gourd (*hue*, Rongorongo sign 124 ✿⁛),[57] meaning a climbing plant that bears fruit". As the whereabouts of Thomson's inscribed calabash are long-lost,[58] and the word-based values allocated by Barthel do not generally replicate over the corpus, we cannot take either of the cited sources beyond a heuristic platform.

All these circumstantial premises, strewn prima facie among the many Easter Island-related publications, suggest that "EISA" may have been conceived as some kind of private device with painted signs working out as a propitious omen. Whether its *mana* was channeled toward a good harvesting or the multiplication of chickens, we claim no explicit authority on the matter. But in each of the cases or in another background that involves pre-modern religious practices, the hypothesis deserves close attention. The "amulets" in use in ancient societies and communities (as the paper label of "EISA" also suggests) are described as fulfilling two broad functions: "apotropaic," meaning the warding off (of evil forces), and "talismanic"-like, i.e., by imbuing or "charming" the bearer / wearer with favor and fortune (Kotansky, 2019, p. 507,

---

56. Barthel (1958, p. 325).

57. The outward shape of glyph /124/ is more likely to depict the *poporo* (*Solanum nigrum*) berries; cf. Métraux (1940, p. 160). The accompanying images do not appear in Barthel's (1978) work; they are extracted from his *Grundlagen…* (1958), and inserted by the current authors for the sake of clarity.

58. We considered and dismissed the idea that Thomson's inscribed calabash might actually be the "EISA" artifact as we believe that Thomson would surely have been able to distinguish between a genuine calabash and a wooden object that is vaguely gourd-shaped.

note 1). The second function in Kotansky (2019) appears to fit conceptually with the recently located artifact from Easter Island (= "EISA"), or with the pre-missionary receptacle-like object illustrated in Fig. 5. If "EISA"'s painted message suggests something like a personalized "calendar" (analogous to the neatly arranged "Lunar Calendar" on "Mamari"), this working line requires serious consideration. R. W. Sproat (2010, p. 126) points out that "a calendar clearly represents a kind of list, and the *Mamari* calendar… represented a fairly sophisticated symbolic representation of an algorithm for maintaining the lunar calendar". Lists are reassuring and tangible reminders of matters concerning directly the welfare and functioning of an individual or a group (cf. in a general context, Belknap, 2004). In this picture, M. Hyman (2006, p. 245), in opposition to the idea that listed formulations may be elementary, naïve, and underrepresented linguistic samples, states,

> Yet less clearly linguistic instances of writing—calendars, tables of sines and cosines, architectural plans, recipes for foods and drugs, mathematical formulae, coins and bank-notes, charts for navigation, computer programs— reflect highly sophisticated intellectual activity and serve as indispensable bearers of culture.

Nonetheless, we should keep in mind that listed items have involved (or not) true writing during the recorded history of humankind. While this intriguing topic is beyond this article's focus, we should constrain ourselves to the extent and mechanisms of the classical *rongorongo* script and encourage solutions from an indigenous perspective prior to 1864, or an early post-1864 one.[59] Lists (and "lunar-based calendars," thereof) are formalized cut-outs of a wider and deeper lore / knowledge that they represent, and they come in different formats, materials, sizes, and colors.[60] It seems the "Mamari lunar calendar" and "EISA" fit in part or in whole under these premises.

In attempting to define an acceptable context, we must pay heed to other hints, too. On the condition that the label was indeed made in or around 1885/6, it clearly shows that the collector or purchaser related it to a "Sacred Amulet". As such, rather than an object simply made and /

---

59. Specifically, it may well be that we are dealing in "EISA"'s case with an object used for propitiation and/or divination in accordance with practical and religious-like scenarios fitting the spirit of the time. For those letting their imaginations run wild, the "encoded message" could have expressed, e.g., the escape velocity of a body from the gravity of the Earth and its trajectory to the Moon. Depending on one's bias- and fantasy-level, other interpretations are likely to be found, see especially Lee's (1999) "The Nutcase Chronicles".

60. "Lists consist of arrangements of entries and have been used for varied purposes throughout history. Lists enumerate, account, remind, memorialize, order. Lists take a number of sizes, shapes and functions, ranging from directories and historical records to edicts and instructions" (Belknap, 2004, p. 6).

or decorated for trade,[61] it seems that "EISA" owed allegiance to its basic function. The hair tuft and bones tied to it are not something, for instance, that we would expect of an object typically made for trade; rather, they are fitting of a genuine amulet (perhaps imparting their *mana* to the piece). The understanding is that items made explicitly for exchange tended to be copies of the conventional wooden statuettes (*mo'ai kavakava*, *mo'ai tangata*, *mo'ai moko*, *mo'ai tangata manu*, and so forth; cf. Geiseler, 1995, p. 66; Balfour, 1917, pp. 359–361; Métraux, 1940, pp. 250–251; Fischer, 2005, p. 75; Hooper, 2006, pp. 145–147, Wieczorek, 2016b, p. 13). However, we can speculate that items (including "sacred objects") made for other purposes initially were in fact traded, especially after the introduction of Christianity—at which time they may have lost their meaning and licit "sacredness," in any case for some Rapanui in need.

4.2.   Pre- or Post-missionary Provenance?

Much of what is considered "rational" in our current fact-gathering and subsequent analysis, must rely on the authenticity of the painted relic,[62] the interaction of the studied variables, and the credibility of consulted bibliographic sources. Since the variables should be supported, it is necessary to direct our attention upon another one, namely, the time-frame of the artifact. The insights gained are relevant to our discussion. It is also worth noting that, up to this point, no single interpretation of *rongorongo* sequences is widely agreed upon by the international researchers, with the exception, perhaps, of the "Lunar Calendar" on "Mamari". The body of speculations is rich enough to fuel entire multi-volume series. Yet, the misleading theories and/or decipherments help us acknowledge the cognitive limitations in facing unknown objects and phenomena. They, similarly, point to the design of a step-by-step process, where any small piece of hard evidence (among the many letdowns) may aid to a broader envisioning and understanding. The time-frame is one of those pieces of evidence that could convey a sounder interpretation of "EISA". Regarding the age of "EISA," close inspection of the paper label instills confidence that it was attached to the object by either the original collector of "EISA," or by an early subsequent owner. We have no reason to doubt the verisimilitude of the statement on the label, namely that the object came from Easter Island in 1885 or 1886. And more to the point, "EISA" is no later in date than 1885/6. Arguably, it might have

---

61. See in a general context, Fischer (1997, p. 509).

62. In formal terms, providing that "EISA"'s original post-1864 function was firmly rooted in the earlier cultural substratum, and was not deliberately made for trade or sale.

been made earlier than its collection date, but based on careful inspection of the object, we doubt that it was created earlier than perhaps a quarter century prior to the time of collection. The wooden object that forms the "body" of "EISA" may have been some sort of discarded European object, and the paint on "EISA" has the appearance of what, in the vernacular of modern collectors, is referred to as "European trade paint". Based on his experience studying various Easter Island objects found in a number of collections (private and public), the "instinct" of RMS is that "EISA" in its present (final) form most likely dates to the period roughly spanning the 1860s to 1885/6.

Assuming that "EISA" took its final form, as we see it today, shortly before being collected in 1885/6, this would place it in the early post-missionary period (circa 1870s to mid-1880s), just prior to the annexation of Easter Island by Chile in 1888. Is the *rongorongo* "inscription" painted on its surface compatible with such a time-frame? Here it is worth noting S. R. Fischer's (1997, p. 10) perceptive comments regarding the possible survival of knowledge during this turbulent time in the island's history:

> [...] it is probable that most *rongorongo* experts and some *rongorongo* pupils actually escaped the devastating 1862–63 labour raids. Further, it is even possible that a few *rongorongo* experts and pupils survived the pandemics that followed. Notwithstanding, what knowledge of script type and function was preserved after this in the 1870s and 1880s constituted a minimal bequest which the next generation wasted through loss, contamination, and invention.

Thus, Fischer suggests that some of the *rongorongo* experts survived with their knowledge into the 1870s and 1880s; "EISA" may be a product of the survival of such knowledge during this period. It is plausible to assume that either the person behind "EISA" had some degree of familiarity with RR, or s/he copied (remembered) signs—especially in consideration of # /152/—from somewhere. This should not be a surprise, as we think that a number of indigenous people still equated the full moon (*O*)*motohi* with the visualization of the "Old woman / man" cooking in the "earth-oven," as Sebastian Englert reports (1948, p. 165). Furthermore, the idea of objects imbued with "supernatural power" for obtaining favors finds another candidate in one of the inherited tablets, the one known as the "Échancrée,"

> The *Échancrée*, probably because of its propitiatory virtues, was summarily transformed into a fishing-line spool before being enveloped in long tresses of sacred hair that were given to Mgr Tepano (figure 197).
> (Orliac and Orliac, 2008, p. 248; cf. Fischer, 1997, p. 22)

This post-1864 act implies that a larger number of fish could have been caught given the "magic power" that permeated this recycled

tablet.[63] Whether in the case of "EISA" or "Échancrée" (or of other objects and personages, for that matter),[64] we should consider how they follow a pattern. These coincidences make us think there is more than meets the eye: magic thinking was ever-present and holding a grip in ancient Rapa Nui, and though in rapid decline,[65] it was still in occasional use in the early post-missionary years.

4.3.   A Plausible Classification?

The last area of discussion pertains to the classification of artifacts such as the "EISA". An artifact may be ethnographically interesting, or it may be epigraphically interesting, or both, or neither. Because matters of ethnographic provenance and epigraphic insight are each focuses of the article, the table[66] below organizes the various possible scenarios.

There are a number of categories and combinations that can be considered here. With respect to ethnographic artifacts there are various possibilities. The first one is pre-tourism artifacts that were manufactured on pre-missionary Easter Island before the time of mass contact with Europeans / North Americans (that is, before circa the mid 1860s). These artifacts are valuable from the point of view of recording the Island's cultural history. Another category is artifacts that were manufac-

---

63. See also W. Hough's (1889, p. 886) short description of charm-stones, "Rude, unshapen (= unshaped) stones were distinguished by the natives as gods of three varieties. These are the fish god in general, called Mea Ika; the bonito's god, called Mea Kahi; and the fowl god, called Mea Moa. The gods were never common, and were possessed by clans or communities, and never by individuals. They were moved about from place to place as they were needed"; W. J. Thomson's (1891, p. 470) subsection "Superstitions," "Fishhooks were made of bones of deceased fishermen, which were thought to exert a mysterious influence over the denizens of the deep. Fishermen were always provided with the stone god that was supposed to be emblematic of the spirit having cognizance of the fish," and J. Macmillan Brown's (1979, pp. 190–191) subsection "Sorcery and Fishing" regarding the use of *mana* charm-stones to secure a successful catch. Otherwise, J. Golson (1965, pp. 62–69), enlists the one-piece and two-pieces fish-hooks (made of stone and bone) recovered from different sites on Easter Island (cf. Heyerdahl and Ferdon, 1961), whereas William Ayres (1979) reports on the fishing techniques and implements used traditionally on Easter Island.

64. See W. Hough (1889, pp. 885–886); W. J. Thomson (1891, p. 470); O. M. Dalton (1904, p. 6); W. Knoche (1914, p. 346); J. Macmillan Brown (1979, pp. 126–128, 134); S. Englert (1948, pp. 267–268); A. Métraux (1957, pp. 88–90, 125–127, 139–143); S. R. Fischer (1997, pp. 331–332); C. Mordo (2002, pp. 73–74); and Harris and Melka (2011b, pp. 264–265).

65. See, e.g., Métraux (1957, p. 127), "The disappearance of a large proportion of the priesthood during the slave-raid of 1862 would explain this sharp break in religious tradition and forgetfulness of the ancient cults".

66. This table is modified from one that Gordon Berthin originally suggested to us; he deserves credit for its inception, but is not responsible for our version of the table.

tured during missionary or post-missionary times as tourist goods, in which case, they are often (but not always) of little ethnographic value when it comes to the study of pre-contact or pre-missionary Rapa Nui; however, they may be of use in comparative tests with items confirmed by scholarship as authentically ancient.

With respect to epigraphic artifacts bearing *rongorongo* or *rongorongo*-like inscriptions or signs, the most sought-after and valued are artifacts that can be demonstrated to have been made during pre-tourism / pre-missionary times inscribed with previously unknown glyph sequences (generally the longer, the more desirable); however, pre-missionary artifacts with glyph sequences that are similar or parallel to the known corpus are highly desirable and valuable epigraphically as well. Also of value epigraphically are artifacts from Easter Island bearing *rongorongo* or *rongorongo*-like inscriptions or signs that are post-missionary, but demonstrate on the part of the creator some genuine familiarity with *rongorongo* traditions, even if minimal or rudimentary, or which are copies of genuine pre-missionary *rongorongo* inscriptions that have since been lost (in analogy, there are no known copies of Plato's dialogues written in his own hand, but rather primarily copies that post-date Plato by over a millennium).

Placing a specific artifact into one of these categories is not always easy. For instance, whether or not "EISA" is a pre-tourist / pre-missionary item is unclear. Even though it was collected in post-missionary times, it is not impossible that it was manufactured in the early 1860s (or earlier) and thus is pre-missionary; however, this has yet to be demonstrated, and is a question that may be impossible to resolve with current information. We currently favor the "conservative" hypothesis that "EISA" dates to late missionary or early post-missionary times, circa late 1860s to early 1880s. More pertinent to the issues which are the focus of this paper, we have provided evidence that the creator of "EISA" minimally possessed some basic familiarity with the genuine *rongorongo* tradition. Thus, referring to Table 1, we consider "EISA" to be a valuable / heuristic epigraphic artifact.

5. Conclusions

The appearance of an unknown (and apparently puzzling) artifact from a different culture represents a challenge for the human cognitive abilities. The basic incongruity we may face is considering "EISA" or other parochial artifacts in our own twenty-first century terms.

Hampering not only analysis of "EISA" specifically, but a full understanding of *rongorongo* more generally, is the possibility that *rongorongo* is an "early script"; that is, a script in an early developmental stage. The *rongorongo* signs may not have corresponded to a spoken language pho-

TABLE 1. Categorization of newly discovered Rapanui-made artifacts, based on their ethnographic provenance and epigraphic validity

Epigraphic significance?	Pre-tourism provenance? [= pre-missionary]	Artifact utility
No epigraphic interest (does not bear a *rongorongo* inscription or is a modern tourist item, reproduction, artwork, or other object with *rongorongo*-like glyphs in imitation of, or inspired by, those found on genuine pre-missionary pieces as published in the modern literature)	No pre-tourist provenance	Of interest primarily to ethnographers and historians considering post-missionary / post-tourist Rapanui culture and Easter Island history
Valid and original (not previously known) *rongorongo* inscription, glyph, or collection of glyphs, created through and informed by genuine indigenous knowledge of *rongorongo* practices (not through familiarity with the modern literature on *rongorongo*) / Copy of an ancient *rongorongo* inscription that is not known otherwise	No pre-tourist provenance	A valuable / heuristic epigraphic artifact
No epigraphic interest (does not bear a *rongorongo* inscription)	Pre-tourist provenance	A valuable ethnographic artifact
Valid and original (not previously known) *rongorongo* inscription / A copy or a "reinterpretation" or a "paraphrase" of a previously known *rongorongo* inscription / A single *rongorongo* glyph or collection of glyphs	Pre-tourist provenance	Most valuable to both ethnographers and epigraphers

netically, word-by-word or syllable-by-syllable.[67] (Compounding the issue, the exact language that gave rise to the *rongorongo* script is still elusive, although it was presumably an ancient language spoken on Easter Island[68] which ultimately gave rise to the historically known language of the Rapanui; the island was first discovered by Europeans in 1722.) Information in an "early script" is communicated not so much by representing the intricacy of a language in detail and specific one-to-one correlations between the written and oral words, but rather through certain specific words, symbols, and other prompts serving as mnemonic devices; metaphorical allusions; homonymy; and other pictorial and semantic indicators.[69] Incident to the above lines, we must clarify that we are not framing *rongorongo* within a teleological model of script development, with *rongorongo* placed at the beginning and the "alphabet" format being the crowning of progression (cf. Moorhouse, 1946, p. 17; Gelb, 1963, pp. 190–205; Pulgram, 1976, p. 4, Table 1.1). Quite the opposite: as far as present evidence reveals to us, the adopted approach helps in avoiding the pitfalls that have plagued many suggested decipherments.

One possible strategy through which it is possible to deduce or even ascertain the nature of "EISA" is "by trial and error" (cf. Ross, 1940, p. 559), in proportion to its specific terms and context within the general socio-cultural setting of pre-missionary to early post-missionary Rapanui. Further objective revisions are welcome in this framework.

The meaning of "EISA" should be construed by accretion and relative to a particular context, in this case, to a lunar-based one and its firsthand use. The most consistent frame of reference for "EISA" is the "Lunar Calendar" on the "Mamari Tablet". Although we are not fully certain about the reliability of the comparison, the analysis shows that a good number of "EISA" signs are found within the "LC," especially in view of the case-sensitive glyph /152/, the centerpiece of said "calendar". Whether or not the glyphic portions (C*a*6—C*a*9 [= C*r*6—C*r*9] ≈ "EISA") are semantically compatible,[70] this is a question that cannot be answered bluntly. Readers should be informed too, that it is not our desire to multiply at all costs the lunar calendars in the *rongorongo* cor-

---

67. Be that as it may, we are eager to read new full translations from researchers who are inclined to "decipher" the *rongorongo* on a syllabic basis, e.g., of the Lunar Calendar on Tablet "Mamari" (cf. Barthel, 1958; Guy, 2006, pp. 63–65; Ávila Fuentealba, 2007, pp. 82–83, 147), or of the "lunar"-like sequences purportedly found on the "Keiti" Tablet (cf. Ávila Fuentealba, 2007, pp. 80, 87; Wieczorek, 2016a).

68. See especially Fischer (2013).

69. See, e.g., Sampson (1985, p. 38) for an assessment of such scripts, "Likewise, the relatively extreme incompleteness of some early scripts may not always be merely a flaw of immaturity; if a script is used only for highly specific purposes, so that much of any utterance is predictable from context…".

70. The adverb "semantically" is resorted to here in face of a number of absent criteria for qualifying the message on "EISA" as fully-fledged phonetic writing.

pus. Any claim regarding the glyphic contents of "EISA" is legitimate only under certain conditions: being the foremost ones, we acknowledge again, the authenticity of the artifact and the commonality of the "full moon" glyph /152/ with that of the "Mamari Tablet". The appearance of new material—i.e., "EISA," or any other potential artifact—raises hopes, however, for replicable research.

As for the core function of "EISA," one can speculate on the basis of the glyphic content and its general physical make-up. The painted "text" seems to be neither informative nor descriptive in nature. At its most basic level, the textual end seems performative: coaxing the artifact so as to achieve the desired outcome. In this vein, "EISA" might have been a personal appliance intended to propitiate the original owner on sustenance practices, or to exercise "divination" on matters of earthly existence.[71] Given the meager living conditions, "astronomy" and "magic"—as understood and ritualized by the Old Rapanui—were in high regard for the express purpose (among others) of achieving effects such as avoiding harmful influences, securing the multiplication of chickens, fishes, turtles, sea birds, or the protection / abundance of harvested plants. Performative writing is attested in different eras and geographical locations, and appears related to basic human needs, emotions, and instincts (see, e.g., Austin, 1962; Tambiah, 1968; Bodel, 2001, pp. 19–24; Page, 2004; Sproat, 2013, pp. 20, 38; Kotansky, 2019).

In light of the collected corpus, and the new pieces that are surfacing (or resurfacing),[72] it becomes clear that there is no such thing as a standard-issue *rongorongo* tablet or format, rather than a mixture of objects and supports intentionally or opportunistically chosen (or salvaged) to be inscribed, painted, scratched, and punched under distinct circumstances and for different purposes (cf. Melka, 2017). The latest artifacts (plus other potential ones which may be discovered in the future) can only be a part of the *rongorongo* story. Yet, taken as a whole they expand the opportunity to revisit and better understand the *rongorongo* practices. An extensive line of authors have commented upon the sacred and hierarchical scale of the tradition (e.g., Thomson, 1891, p. 514; Dalton, 1904, p. 6; Routledge, 1919, pp. 245–246; Brown, 1979, p. 74; Métraux, 1940, p. 395; Englert, 1948, p. 316; McCoy, 1979, p. 158; Fischer, 1997, p. 555). The media and formats used, plus the modifications, abridgments, and "re-editions" of "texts" through time, suggest however an activity beyond the core elite of the *rongorongo* scribes. Perhaps it is appropriate to reach M. de Laat's (2009, p. 219) thought,

---

71. Since Old Easter Islanders had "…numerous superstitions and resorted to charms, incantations… and amulets… (Thomson, 1891, p. 469)," the artifact known as "EISA," or the objects illustrated and commented upon in footnote 2 (v. *supra*) and Figs. 5 and 6 for that matter, would not be out of favor.

72. See Melka and Schoch (2020a,b) and Schoch and Melka (2019; 2020b).

The fact that, at the time of the first mentioning by Eyraud in 1864, tablets were present in all the huts… also poses the intriguing problem how widely at one time literacy had spread beyond the cultural elite.

Several objects with fake RR signs have received a good deal of attention in the published literature.[73] A careful description and elucidation of "EISA" perhaps will assist in identifying what is *phony* or *half true* in a domain characterized by so much wishful thinking and where scholarly opinions may not sit easily together. Of course where one draws the boundary between *a fake post-1864 item* and *an authentic post-1864 item* depends upon the impartial analysis of the hitherto amassed evidence (at best), or acting on personal assumptions (at worst). Any misinterpretation may be especially bound to happen, due, for example, to the insufficiency of data about the links of the long chain of entities involved in "EISA"'s ownership: (1) the original indigenous painter; (2) the purported European collector / purchaser, i.e., the "Irish missionary"; (3) any potential subsequent owner (?); (4) the English anthropologist and collector Harry Geoffrey Beasley; (5) after HGB's decease in February 1939, any potential subsequent private owner (?) / or an institution (?); (6) the next anonymous English collector who sold the piece in 1985; (7) the Hawaiian buyer who traded "EISA" later with (8) the Canada-based antiques dealer.

Although most likely a post-missionary *rongorongo* product and standing for some sort of "pocket calendar" (to the best of our assessment), "EISA" is scientifically desirable in its own right. On the face of it, one would not expect that all knowledge of *rongorongo* would be absolutely and completely lost as the result of the Peruvian labor raids of 1862–1863, the coming of the first documented missionary in 1864, and the later missionary work (cf. Eyraud, 1866; Fischer, 1997, pp. 9–10). In this sense, "EISA"—collected during post-missionary times—records a personal effort to continue with the *rongorongo* tradition in approximately the last third of the nineteenth century.

The present authors would gladly agree to the further expansion of the corpus (whether related to the work of the original painter of the "EISA" glyphs or otherwise). The fixed-content "text" of "EISA," scattered and short as it appears, is an accidentally preserved trace of an unknown pre-twentieth-century Rapanui individual. Any comparison with similarly painted / written texts would have increased the chances to study stylistic features, e.g., morphological variation; to explore whether these kinds of painted "amulets" (in gourd-like shapes or not) were casually or systematically manufactured; to examine the material support of the new "texts," et cetera. By the same token, valuable

---

73. Cf. Métraux (1940); Imbelloni (1951); Barthel (1958); van Hoorebeeck (1979); Forment and Esen-Baur (1990); Fischer (1993); Schoch and Melka (2020b).

and distinctive as it historically is, the very concise "1885/6" note is not telling much about the identity of the first purchaser / collector, or the circumstances of "EISA"'s acquisition. In either case, we are conditioned in our search by what is physically accessible. Yet, in order to sustain the hope for further scientific investigation, one may wonder if additional genuine RR-inscribed pieces are still lying dormant somewhere among private and museum collections waiting to be discovered and evaluated.

Until recently, the *rongorongo* corpus has been relatively static, with the known and "accepted" texts limited to just over two dozen items (Barthel, 1958; Fischer, 1997). Our research has included bringing additional pieces from Easter Island bearing *rongorongo* signs and sequences from the late pre-missionary to early post-missionary period, circa 1860s to 1880s, to the attention of interested scholars. In addition to the "Sacred Amulet from Easter Island," described herein, we have documented the "*Rangitoki* bark-cloth fragment" (Fig. 7 above, collected on Easter Island in March 1869; Schoch and Melka, 2019; Schoch and Melka, 2020b) and the "*San Diego* Tablet" (Fig. 8 above, possibly dating to the circa late 1850s − early 1860s or shortly thereafter; Melka and Schoch, 2020a). Here we wish to express the conviction and hope that these "newly unveiled" artifacts, and, possibly, future items that may come to light, will aid researchers in their studies of the *rongorongo* script.

Acknowledgments

References

Äkerblom, Kjell (1968). *Astronomy and Navigation in Polynesia and Micronesia*. Vol. 14. Monograph Series. Stockholm: Etnografiska Museet.

Albright, Daniele (2000). *Untwisting the Serpent: Modernism in Music, Literature and Other Arts*. Chicago: The University of Chicago Press.

Anonymous (2005a). "CEIPP [Cercle d'Études sur l'Ile de Pâques et la Polynésie]." In: *The Rongorongo of Easter Island. Item C: The Mamari Tablet*. URL: http://kohaumotu.org/rongorongo_org/corpus/mamari.html (visited on 24 April 2019).

Anonymous (2005b). "CEIPP [Cercle d'Études sur l'Ile de Pâques et la Polynésie]." In: *The Rongorongo of Easter Island: The Lunar Calendar on Tablet C (Mamari)*. URL: http://kohaumotu.org/rongorongo_org/rosetta/lunar.html (visited on 24 April 2019).

Austin, John Langshaw (1962). *How to Do Things with Words*. Oxford: Clarendon Press.

Ávila Fuentealba, Franklin (2007). "Ensayo de Estudio Visual de las Tablillas Roŋoroŋo." Temuco, Chile. URL: http://www.serindigena.org/archivosdigitales/otros/ensayo_rongorongo_de_tocariov.pdf (visited on 1 February 2020).

Ayres, William S. (1979). "Easter Island Fishing." In: *Asian Perspectives* 22.1, pp. 61–92.

Ayres, William S., Becky Saleeby, and Candace B. Levy (2000). "Late Prehistoric-Early Historic Easter Island Subsistence Patterns." In: *Easter Island Archaeology: Research on Early Rapanui Culture*. Ed. by Christopher M. Stevenson and William S. Ayers. Los Osos, California: The Easter Island Foundation / Bearsville Press, pp. 191–204.

Baayen, Harald R. (2001). *Word Frequency Distributions*. Ed. by Nancy Ide and Jean Véronis. Vol. 18. Text, Speech and Language Technology. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Balfour, Henry (1917). "Some Ethnological Suggestions in regard to Easter Island, or Rapanui." In: *Folklore* 28.4, pp. 356–381.

Bantock, Nick (2000). *The Artful Dodger: Images and Reflections*. San Francisco: Chronicle Books.

Baroni, Marco (2006). "Distributions in Text." In: *Counting Words: An Introduction to Lexical Statistics. Introductory Course, Language and Computation Section. The 18th European Summer School in Logic, Language, and Information, 31 July–11 August, 2006– Málaga, Spain (ESSLLI 2006)*. Ed. by Marco Baroni and Stephan Evert, pp. 1–22. URL: http://sslmit.unibo.it/ (visited on 6 May 2019).

Barthel, Thomas S. (1958). *Grundlagen zur Entzifferung der Osterinselschrift*. Vol. 36. Reihe B. Völkerkunde, Kulturgeschichte und Sprachen. Hamburg: Cram, de Gruyter & Co.

————— (1963). "*Rongorongo*-Studien (Forschungen und Fortschritte bei der weiteren Entzifferung der Osterinselschrift)." In: *Anthropos* 58.3–4, pp. 372–436.

————— (1972). "Zur Frage der lunaren Zeichen in der Osterinselschrift." In: *Asian and African Studies* 8, pp. 9–18.

————— (1978). *The Eighth Land: The Polynesian Discovery and Settlement of Easter Island*. Trans. by Anneliese Martin. Honolulu: The University Press of Hawai'i.

————— (1993). "Perspectives and Directions of the Classical Rapanui Script." In: *Easter Island Studies. Contributions to the History of Rapanui in Memory of William T. Mulloy*. Ed. by Steven R. Fischer. Vol. 32. Oxbow Monographs. Oxford, UK: Oxbow Books, pp. 174–176.

Batista Campbell, Ramón (1971). *La Herencia Musical de Rapanui: Etnomusi- cología de la Isla de Pascua [The Musical Legacy of Rapanui: Ethnomusicology of Easter Island]*. Santiago de Chile: Editorial Andrés Bello.

Beasley, Harry Geoffrey (1928). *Pacific Island Records: Fish Hooks*. London: Seeley, Service & Co.

Belknap, Robert E. (2004). *The List: The Uses and Pleasures of Cataloguing*. New Haven, London: Yale University Press.

Belmonte Avilés, Juan Antonio (2008). *Tiempo y Religión: Una Historia Sagrada del Calendario [Time and Religion: A Sacred History of the Calendar]*. Vol. 22. Colección Religiones y Textos. Biblioteca de las Religiones. Madrid: Ediciones Clásicas · Ediciones Del Orto.

Berthin, Gordon and Michael Berthin (2006). "Astronomical Utility and Poetic Metaphor in the Rongorongo Lunar Calendar." In: *Applied Semi- otics / Sémiotique appliquée* 18.12, pp. 85–99.

Bianco, Jean (1976). "Thomas Barthel et le déchiffrement de l'écriture pascuane (1re partie)." In: *Kadath* 20, pp. 13–21.

Biber, Douglas (1993). "Representativeness in Corpus Design." In: *Liter- ary and Linguistic Computing* 8.4, pp. 243–257.

Bodel, John (2006). "Epigraphy and the Ancient Historian." In: *Epi- graphic Evidence: Ancient History from Inscriptions*. Ed. by John Bodel. Lon- don & New York: Routledge, pp. 1–56.

Boone, Elizabeth Hill (2009). "When Art is Writing and Writing, Art: Graphic Communication in Preconquest Mexico." In: *Dialogues in Art History, from Mesopotamian to Modern: Readings for a New Century*. Ed. by Elizabeth Cropper. Vol. 74. Studies in the History of Art. New Haven and London: National Gallery of Art, Washington, pp. 57–73.

Brookman, David Y. (2007). "Easter Island Home Page. Easter Is- land's Rongorongo Script." see `https : / / web . archive . org / web / 20070830125104/http://www.netaxs.com/~trance/mamari.html` (accessed 29 May 2020) and `https://web.archive.org/web/20070810225655/http: //www.netaxs.com/~trance/rapanui.html` (accessed 29 May 2020). URL: `http://www.netaxs.com` (visited on 21 February 2008).

Brown, John Macmillan (1979). *The Riddle of the Pacific*. Reprint of the 1924 edition. Adelphi Terrace, London: T. Fisher Unwin Ltd. New York: AMS.

Buck, Peter H. [Te Rangi Hiroa] (1938). *Vikings of the Sunrise*. New York: Frederick A. Stokes.

Butinov, Nikolai A. and Yuri V. Knorozov (1957). "Preliminary Report on the Study of the Written Language of Easter Island." In: *Journal of the Polynesian Society* 66.1, pp. 5–17.

Carreau, Lucie (2010). "Becoming 'Professional': From the Beasley Col- lection to the Cranmore Ethnographical Museum." In: *Journal of Mu- seum Ethnography* 23, pp. 41–55.

Chadwick, John (2000). *The Decipherment of Linear B*. Cambridge: The Press Syndicate of the Cambridge University.

Chauvet, Stéphen-Charles (1935). *L'Île de Pâques et Ses Mystères*. Trans. by Ann M. Altman. Ed. by Shawn McLaughlin (2005). Paris: TEL. URL: http://www.chauvet-translation.com/index.htm (visited on 18 January 2021).

Churchill, William (1912). *The Rapanui Speech and the Peopling of Southeast Polynesia, Publication No. 174*. Washington: The Carnegie Institution of Washington.

Coe, Michael and Justin Kerr (1997). *The Art of the Maya Scribe*. London: Thames and Hudson.

Cooke, George H. (1899). "Te Pito te Henua, known as Rapa-Nui, commonly called Easter Island, South Pacific Ocean." In: *Annual Reports of the Smithsonian Institution for 1897*. Washington: Smithsonian Institution, United States National Museum, pp. 689–723.

Corliss, William R. (2005). *Archaeological Anomalies: Graphic Artifacts I Coins, Calendars, Geoforms, Maps, Quipus*. Glen Arm, MD: The Sourcebook Project.

Craig, Robert D. (2004). *Handbook of Polynesian Mythology*. Santa Barbara, California: ABC–CLIO, Inc.

Dalton, Ormonde Maddock (1904). "On an Inscribed Wooden Tablet from Easter Island (Rapa Nui) in the British Museum." In: *Man* 4, pp. 1–7.

Damerow, Peter (1996). *Abstraction and Representation: Essays on the Cultural Evolution of Thinking*. Trans. by Renate Hanauer. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Daniels, Peter T. and William Bright, eds. (1996). *The World's Writing Systems*. New York and Oxford: Oxford University Press.

Davis, Tom (2007). "The Practice of Handwriting Identification." In: *The Library: The Transactions of the Bibliographical Society* 8.3, pp. 251–276.

de Laat, M. (2009). *Words out of Wood: Proposals for the Decipherment of the Easter Island Script*. Delft, The Netherlands: Eburon Academic.

Dederen, François and Steven R. Fischer (1993). "The Traditional Production of the Rapanui Tablets." In: *Easter Island Studies: Contributions to the History of Rapanui in Memory of William T. Mulloy*. Ed. by Steven R. Fischer. Vol. 32. Oxbow Monographs. Oxford, UK: Oxbow Books.

DeFrancis, John (1989). *Visible Speech: The Diverse Oneness of Writing Systems*. Honolulu: University of Hawai'i Press.

Eggertsson, Sveinn (2011). "Human Figures in Rapanui Woodcarving." In: *Journal of the Polynesian Society* 120.2, pp. 113–128.

Elkins, James (1996). "On the Impossibility of Close Reading: The Case of Alexander Marshack." In: *Current Anthropology* 37.2, pp. 185–226.

Englert, Sebastian (1948). *La Tierra de Hotu Matu'a: Historia, Etnología y Lengua de la Isla de Pascua [The Land of Hotu Matu'a: History, Ethnology, and Language of Easter Island]*. Santiago de Chile: Padre las Casas.

———— (1970). *Island at the Center of the World: New Light on Easter Island*. New York: Charles Scribner's Sons.

Eyraud, Joseph-Eugène (1866). "Lettre du F. Eugène Eyraud, de la Congrégation des Sacrés-Cœurs de Jésus et de Marie, au T. R. P. Supérieur général de la même Congrégation à Paris. Valparaiso, décembre 1864." In: *Annales de la Propagation de la Foi—Recueil Périodique des Lettres des Evêques et des Missionnaires des Missions des Deux Mondes, et de tous les Documents Relatifs aux Missions et à l'Œuvre de la Propagation de la Foi* 38, pp. 52–71, 124–138.

Facchetti, Giulio M. (2002). *Antropologia della scrittura: Con un'appendice sulla questione del Rongorongo dell'Isola di Pasqua*. Milan: Arcipelago Edizioni.

Fischer, Steven R. (1993). "A Provisional Inventory of the Inscribed Artifacts in the Three Rapanui Scripts." In: *Easter Island Studies: Contributions to the History of Rapanui in Memory of William T. Mulloy*. Ed. by Steven R. Fischer. Vol. 32. Oxbow Monographs. Oxford, UK: Oxbow Books.

———— (1994). "Rapanui's 'Great Old Words': E Timo Te Akoako." In: *Journal of the Polynesian Society* 103.4, pp. 413–443.

———— (1997). *Rongorongo: The Easter Island Script, History, Traditions, Texts*. Oxford: Oxford University Press.

———— (2005). *Island at the End of the World: The Turbulent History of Easter Island*. London: Reaktion Books.

———— (2013). "Sources of the Old Rapanui Language of Easter Island." In: *Oceanic Voices—European Quills: The Early Documents on and in Chamorro and Rapanui (Colonial and Postcolonial Linguistics)*. Ed. by Steven R. Fischer. Berlin: Academie Verlag, pp. 11–23.

Forment, Francina and Heide-Margaret Esen-Baur, eds. (1990). *L'Île de Pâques: Une énigme? Bruxelles: Musées Royaux d'Art et d'Histoire*. Bruxelles: Musées Royaux d'Art et d'Histoire / Frankfurt am Main: Verlag Philipp von Zabern.

Friedrich, Johannes (1971). *Extinct Languages*. Translated from the German, Entzifferung verschollener Schrifton und Sprachen by Frank Gaynor. Westport, Connecticut: Greenwood Press, Publishers.

Gaur, Albertine (1987). *A History of Writing*. 2nd ed. London: The British Library.

———— (1994). *A History of Calligraphy*. New York: River Press / Abbeville Publishing Group.

Geiseler, Wilhelm (1995). *Die Oster-Insel (Eine Stätte prähistorischer Kultur in der Sudsee). Geiseler's Easter Island Report: An 1880s Anthropological Account. With an Introduction, Annotations and Notes, by William S. Ayres, translated by William S. Ayres and Gabriella S. Ayres*. Vol. 12. Asian and Pacific Archaeology series. Honolulu: University of Hawai'i at Manoa.

Gelb, Ignace J. (1963). *A Study of Writing*. 2nd ed. Chicago, IL: University of Chicago Press.

Gelb, Ignace J. and R. M. Whiting (1975). "Methods of Decipherment." In: *Journal of the Royal Asiatic Society of Great Britain and Ireland* 107, pp. 95–104.

Godart, Louis and Jean-Pierre Olivier (1982). *Recueil des Inscriptions en Linéaire A*. Vol. 4. Paris: P. Geuthner depositaire.

Golson, J. (1965). "Thor Heyerdahl and the Prehistory of Easter Island." In: *Oceania* 36.1, pp. 38–83.

Gossen, Gary H. (1974). "A Chamula Solar Calendar Board from Chiapas, Mexico." In: *Mesoamerican Archaeology: New Approaches*. Ed. by Norman Hammond. Austin: University of Texas Press, pp. 217–253.

Gradín, Carlos J., Carlos A. Aschero, and Ana M. Aguerre (1976). "Investigaciones Arqueológicas en la Cueva de las Manos Estancia Alto Río Pinturas (Provincia de Santa Cruz)." In: *Relaciones de la Sociedad Argentina de Antropología (Buenos Aires)* 10, pp. 201–250.

Grünbaum, Branko and Geoffrey C. Shepherd (1987). *Tilings & Patterns*. 2nd ed. New York: W. H. Freeman & Company.

Gusinde, Martin (1922). "Bibliografía de la Isla de Pascua, Continuación." In: *Publicaciones del Museo de Etnología y Antropología de Chile* 2.3, pp. 261–383.

Guy, Jacques B.M. (1985). "On a Fragment of the 'Tahua' Tablet." In: *Journal of the Polynesian Society* 94, pp. 367–387.

———— (1990). "The Lunar Calendar of Tablet 'Mamari'." In: *Journal de la Société des Océanistes* 91, pp. 135–149.

———— (1999). "Peut-on se fonder sur le témoignage de Metoro pour déchiffrer les *Rongo-Rongo*?" In: *Journal de la Société des Océanistes* 108.1, pp. 125–132.

———— (2006). "General Properties of the *Rongorongo* Writing." In: *Rapa Nui Journal* 20.1, pp. 53–66.

Harris, Martyn and Tomi S. Melka (2011a). "The *Rongorongo* Script: On a Listed Sequence in the *recto* [*verso*, repaired] of Tablet 'Mamari'." In: *Journal of Quantitative Linguistics* 18.2, pp. 122–173.

———— (2011b). "The *Rongorongo* Script: On a Listed Sequence in the *recto* [*verso*, repaired] of Tablet 'Mamari' (Part II)." In: *Journal of Quantitative Linguistics* 18.3, pp. 234–272.

Harris, Roy (1986). *The Origin of Writing*. Duckworth: University of Oxford.

———— (2001). *Rethinking Writing*. London: Continuum.

Heyerdahl, Thor (1965). "The Concept of RONGO-RONGO among the Historic Population of Easter Island. Report 16." In: *Miscellaneous Papers: Reports of the Norwegian Archaeological Expedition to Easter Island and East Pacific*. Ed. by T. Heyerdahl and Edwin N. Ferdon Jr. Vol. 24. Monographs of the School of American Research and the Kon-Tiki Museum. Stockholm: Forum Publishing House, pp. 345–385.

———— (1975). *The Art of Easter Island. Garden City*. New York: Doubleday & Company, Inc.

Heyerdahl, Thor and Edwin N. Ferdon Jr., eds. (1961). *Archaeology of Easter Island: Reports of the Norwegian Archaeological Expedition to Easter Island and*

East Pacific. Monographs of the School of American Research and the Museum of New Mexico. Stockholm: Forum Publishing House.

Hiroa, Rangi (1938). *Ethnology of Mangareva*. Vol. 157. Bishop Museum Bulletin. Honolulu: Bishop Museum Press.

Hooper, Steven (2006). *Pacific Encounters: Art & Divinity in Polynesia 1760–1860. (Published to accompany the exhibition Pacific Encounters: Art & Divinity in Polynesia 1760–1860, Sainsbury Centre for Visual Arts, University of East Anglia, Norwich, 21 May–13 August 2006)*. London: The British Museum Press.

Horley, Paul (2007). "Structural Analysis of *Rongorongo* Inscriptions." In: *Rapa Nui Journal* 21.1, pp. 25–32.

———— (2010). "*Rongorongo* Tablet Keiti." In: *Rapa Nui Journal* 24.1, pp. 45–56.

———— (2011). "Lunar Calendar in *Rongorongo* Texts and Rock Art of Easter Island." In: *Journal de la Société des Océanistes* 132, pp. 17–38.

Horley, Paul, Albert Davletshin, and Rafał M. Wieczorek (2018). "How Many Scripts were there on Easter Island?" In: *The Sleep of Reason Produces Monsters: Misconceptions about Easter Island in the Light of 21st Century Science*. Ed. by Zuzanna Jakubowska-Vorbrich. Warsaw: Museum of the History of the Polish Popular Movement, Institute of Iberian, and Ibero-American Studies, University of Warsaw, pp. 323–468.

Horley, Paul and Georgia Lee (2012). "Easter Island's Birdman Stones in the Collection of the Peabody Museum of Archaeology and Ethnology." In: *Rapa Nui Journal* 26.1, pp. 5–20.

Hough, Walter (1889). "Notes on the Archeology and Ethnology of Easter Island." In: *The American Naturalist* 23.10, pp. 877–888.

Houston, Stephen D. (2004a). "The Archaeology of Communication Technologies." In: *Annual Review of Anthropology* 33, pp. 223–250.

———— (2004b). "Writing in Early Mesoamerica." In: *The First Writing: Script Invention as History and Process*. Ed. by Stephen D. Houston. Cambridge, UK: Cambridge University Press, pp. 274–309.

Hunston, Susanne (2002). *Corpora in Applied Linguistics*. Cambridge, UK: Cambridge University Press.

Hyman, Malcolm D. (2006). "Of Glyphs and Glottography." In: *Language & Communication* 26.3–4, pp. 231–249.

Imbelloni, José (1951). "Las 'Tabletas Parlantes' de Pascua, Monumentos de un Sistema Gráfico Indo-oceánico [The 'Talking Tablets' of Easter Island, Monuments of an Indo-Oceanic Graphic System]." In: *Runa, Archivo para las Ciencias del Hombre (Buenos Aires)* 4.1–2, pp. 89–177.

Jannot, Jean-René (2005). *Religion in Ancient Etruria*. Trans. by Dane K. Whitehead. Wisconsin Studies in Classics. Madison, Wisconsin: University of Wisconsin Press.

Jaussen, Florentine É. (1893). *L'Île de Pâques. Historique, écriture et répertoire des signes des tablettes ou bois d'hibiscus intelligents*. Ed. by Rev. P. Ildefonse Alazard. Paris: Ernest Leraux.

Jones, Owen (1856). *The Grammar of Ornament: Illustrated by Examples from Various Styles of Ornament*. London: Day and Son.

Jurafsky, Daniel and James H. Martin (2018). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd ed. URL: https://web.stanford.edu/~jurafsky/slp3/ed3%20book.pdf (visited on 20 October 2018).

Kaeppler, Adrienne L. (2003). "Sculptures of Barkcloth and Wood from Rapa Nui: Symbolic Continuities and Polynesian Affinities." In: *RES: Anthropology and Aesthetics* 44, pp. 10–69.

Kjellgren, Erik, ed. (2001). *Splendid Isolation: Art of Easter Island*. With contributions by J. A. Van Tilburg and A. L. Kaeppler. New York/New Haven-London: The Metropolitan Museum of Art/Yale University Press.

Knoche, Walter (1914). "Cráneos Marcados de la Isla de Pascua [Incised crania from Easter Island]." In: *Revista Chilena de Historia y Geografía (Santiago)* 12.16, pp. 344–346.

Köhler, Reinhard and Reinhard Rapp (2007). "A Psycholinguistic Application of Synergetic Linguistics." In: *Glottometrics* 15, pp. 62–70.

Kotansky, Roy D. (2019). "Textual Amulets and Writing Traditions in the Ancient World." In: *Guide to the Study of Ancient Magic*. Ed. by David Frankfurter. Leiden / Boston: Koninklijke Brill NV, pp. 507–554.

Krämer, Sybille (2003). "'Schriftbildlichkeit' oder: Über eine (fast) vergessene Dimension der Schrift." In: *Bild, Schrift, Zahl*. Ed. by Sybille Krämer and H. Bredekamp. München: Wilhelm Fink, pp. 157–176.

Krippendorff, Klaus (2004). *Content Analysis: An Introduction to its Methodology*. 2nd ed. Thousand Oaks, California: Sage Publications.

Krupa, Viktor (1971). "'Moon' in the writing of Easter Island." In: *Oceanic Linguistics* 10.1, pp. 1–10.

Kudrjavtsev, Boris G. (1949). "Письменность острова Пасхи [Scripts of Easter Island]." In: *Сборник музея антропологии и этнографии [Collection of the Museum of Anthropology and Ethnography]*. Vol. 11, pp. 175–221.

Le Ny, Jean-François and Françoise Cordier (2004). "Contribution of Word Meaning and Components of Familiarity to Lexical Decision: A Study with Pseudowords Constructed from Words with Known or Unknown Meaning." In: *Current Psychology Letters* 12.1, pp. 1–10.

Lee, Georgia (1999). "The Nutcase Chronicles." In: *South American Explorer Magazine* 55, pp. 14–20.

Lehmann, Walter (1907). "Essai d'une monographie bibliographique sur l'île de Pâques." Trans. by R. P. Théophané Calmes (des Sacrés Cœurs de Picpus). In: *Anthropos* 2, pp. 141–151, 257–268.

Lelièvre, F. et al. (2010). *Easter Island: An Epic Voyage*. Montréal: Pointe-à-Callière.

Lewis, David (1972). *We, the Navigators—Ancient Art of Landfinding*. Foreword by S. H. Riesenberg. Honolulu: University of Hawai'i Press.

Liller, William (1993). *The Ancient Solar Observatories of Rapanui: The Archaeoastronomy of Easter Island (The Easter Island Series)*. Woodland, California: Easter Island Foundation / Cloud Mountain Press.

Macri, Martha J. (1996). "*RongoRongo* of Easter Island." In: *The World's Writing Systems*. Ed. by Peter T. Daniels and William Bright. Oxford, NY: Oxford University Press, pp. 183–188.

Magli, Giulio (2009). *Mysteries and Discoveries of Archaeoastronomy: From Giza to Easter Island*. New York: Copernicus Books / Springer Science + Business Media: in association with Praxis Pub.

Maiani, Margherita and Soizic Quer, eds. (1996). *Voyage vers l'île mystérieuse. De la Polynésie à l'île de Pâques, 20 Avril–15 Septembre 1996. Musée d'Aquitaine, Bordeaux, Mairie de Bordeaux*. Trans. by Egle Barone Visigalli and Olivier Puaux. Milan: Amilcare Pizzi Editore.

Massarelli, Ricardo (2014). *I Testi Etruschi su Piombo [The Etruscan Texts (appearing) on Lead]*. Vol. 53. Biblioteca di Studi Etruschi. Pisa / Roma: Fabrizio Serra editore.

Mazière, Francis (1968). *Mysteries of Easter Island*. Trans. by Wm. Collins Sons. New York: W. W. Norton, Company, Inc. With photographs by the author. Originally published as *Fantastique île de Pâques: Des yeux regardent les étoiles...*, Robert Laffont, 1965.

McCall, Grant (2010). "The End of the World at the End of the Earth: Retrospective Eschatology on Rapanui (Easter Island)." In: *Online Proceedings of the Symposium "Anthropology and the Ends of Worlds"*. Ed. by Sebastian Job and Linda Connor. Vol. 1. Sydney Anthropology Symposium Series. Sydney: University of Sydney, pp. 45–53.

McCoy, Patrick (1979). "Easter Island." In: *The Prehistory of Polynesia*. Ed. by Jesse D. Jennings. Cambridge, MA / London, England: Harvard University Press, pp. 135–147.

McEnery, Tony and Anthony Wilson (2001). *Corpus Linguistics*. 2nd ed. Edinburgh Textbooks in Empirical Linguistics. Edinburgh, UK: Edinburgh University Press.

Meinicke, Carl Eduard (1875). *Die Inseln des Stillen Oceans, eine geographische Monographie. Zweiter Theil. Polynesien und Mikronesien*. Leipzig: Frohberg.

Melka, Tomi S. (2007). "Structural and Distributional Analysis of the *rongorongo* Text in Tablet 'Mamari'." Manuscript in the collection of TSM.

———— (2009). "The Corpus Problem in the *rongorongo* Studies." In: *Glottotheory: International Journal of Linguistics* 2.1, pp. 111–136.

———— (2013). "'Harmonic'-like structures in the *rongorongo* script." In: *Glottotheory: International Journal of Linguistics* 4.2, pp. 115–139.

———— (2014). "Palindrome-like Structures in the *rongorongo* Script." In: *Empirical Approaches to Text and Language Analysis, Dedicated to Luděk Hřebíček on the Occasion of his 80th Birthday*. Ed. by Gabriel Altmann et al. Vol. 17. Studies in Quantitative Linguistics. Lüdenscheid, Germany: RAM-Verlag, pp. 153–181.

Melka, Tomi S. (2016). "Distinctive Sequences in *rongorongo* Texts." In: *Easter Island: Cultural and Historical Perspectives*. Ed. by Ian Conrich and Hermann Mückler. Berlin: Frank & Timme, pp. 217–236.

———— (2017). "A Developmental Continuum for the *rongorongo* Script of Easter Island. Part I." Manuscript in the Collection of TSM.

Melka, Tomi S. and Robert M. Schoch (2020a). "Exploring a Mysterious Tablet from Easter Island: The Issues of Authenticity and Falsifiability in *rongorongo* Studies." In: *Cryptologia* 44.6, pp. 481–544.

———— (2020b). "The Quest for Information Retrieval: An Inscribed Relic from Ancient Rapa Nui (Easter Island)." Manuscript in the Collections of RMS and TSM.

Meller, Harald (2007). "The Nebra Sky Disc: The Oldest Representation of Heavens." In: *Discovery! Unearthing the New Treasures of Archaeology*. Ed. by Brian Fagan. London: Thames & Hudson, pp. 188–189.

Métraux, Alfred (1940). *Ethnology of Easter Island*. Vol. 160. Bishop Museum Bulletin. Honolulu: Bernice P. Bishop Museum Press.

———— (1957). *Easter Island: A Stone-age Civilization of the Pacific*. Trans. by Michael Bullock. New York: Oxford University Press.

Michelot, Paul and Jean-Claude Michelot (1979). *L'Ile de Pâques démystifiée*. Paris: Librairie Académique Perrin.

Moorhouse, Alfred C. (1946). *Writing and the Alphabet*. Vol. III. Past and Present Studies in the History of Civilisation. London: Cobbett Press.

Mordo, Carlos (2002). *Easter Island*. Trans. by Graciela Smith. Auckland, New Zealand: Firefly Books / David Bateman Ltd.

NMNH, Smithsonian (2014). *Gourd Vessels*. Department of Anthropology Collections. Cat. No. E129756, E129757, E129758. Washington, DC: National Museum of Natural History, Smithsonian. URL: http://collections.nmnh.si.edu/search/anth/ (visited on 20 June 2016).

Ojeda, Carlos Charlin (1947). *Geo-etimología de la ISLA de PASCUA [Geological etymology of EASTER ISLAND]*. Santiago de Chile: Instituto Geográfico Militar.

Orliac, Michel and Catherine Orliac (2008). *Trésors de l'île de Pâques / Treasures of Easter Island*. Collection de la Congrégation des Sacrés-Cœurs de Jésus et de Marie SS. CC. Genève: Frédéric Dawance / Paris: Louise Leiris.

Page, Sophie (2004). *Magic in Medieval Manuscripts*. Toronto and Buffalo: University of Toronto Press.

Palmer, John Linton (1876). "On Some Tablets Found in Easter Island." In: *Proceedings of the Literary and Philosophical Society of Liverpool*. Vol. 30, pp. 255–263.

Polo Müller, Regina (1992). "Mensagens visuais na ornamentação corporal Xavante [Visual messages in the body decoration of Xavante]." In: *Grafismo Indígena: Estudos en Antropologia Estética*. Ed. by Lux Vidal et al. São Paulo: Studio Nobel / Editora da Universidade de São Paulo / FAPESP, pp. 133–143.

Pozdniakov, Konstantin (1996). "Les bases du déchiffrement de l'écriture de l'île de Pâques." In: *Journal de la Société des Océanistes* 103.2, pp. 289–303.

Pulgram, Ernst (1976). "The Typologies of Writing Systems." In: *Writing without Letters*. Ed. by William Haas. Vol. 4. Mont Follick Series. Manchester: Manchester University Press / Totowa, New Jersey: Rowman & Littlefield, pp. 1–29.

Roberts, Mere, Frank Weko, and Liliana Clarke (2006). *Maramataka: The Māori Moon Calendar*. Research Report 283. Matauranga Māori and Bio Protection Research Team National Centre for Advanced Bio-Protection Technologies. Canterbury, NZ: Lincoln University.

Robinson, Andrew (2002). *Lost Languages: The Enigma of the World's Undeciphered Scripts*. London: BVA / New York: McGraw-Hill.

Rogers, Henry (2005). *Writing Systems: A Linguistic Approach*. Malden, MA / Oxford, UK: Blackwell Publishing.

Ross, Alan S.C. (1940). "The Easter Island Tablet Atua-Mata-Riri." In: *Journal of the Polynesian Society* 49.196, pp. 556–563.

Routledge, Katherine (1919). *The Mystery of Easter Island. The Story of an Expedition*. London and Aylesbury: Hazell, Watson and Viney, LD.

Sampson, Geoffrey (1985). *Writing Systems: A Linguistic Introduction*. London: Hutchinson & Co.

Schmandt-Besserat, Denise (1994). "Forerunners of Writing: The Social Implications." In: *Writing Systems and Cognition: Perspectives from Psychology, Physiology, Linguistics, and Semiotics*. Ed. by W.C. Watt. Vol. 6. Neuropsychology and Cognition. Dordrecht / Boston / London: Kluwer Academic Publishers, pp. 303–310.

Schoch, Robert M. and Tomi S. Melka (2019). "The *Raŋitoki* (Rangitoki) Bark-cloth Piece: A Newly Recognized *rongorongo* Fragment from Easter Island." In: *Asian and African Studies (Bratislava)* 28.2, pp. 113–148 and 413–417.

———— (2020a). "A 'Sacred Amulet from Easter Island—1885/6—': Analyzing Enigmatic Glyphic Characters in the Context of the *rongorongo* Script." Poster session presented on 17 June 2020 at the conference Grapholinguistics in the 21st Century (`https://grafematik2020.sciencesconf.org/`). Poster of eight pages available for download from `http://www.fluxus-editions.fr/grafematik2020-files/schoch-slides.pdf` (visited on 27 June 2020).

———— (2020b). "The *Raŋitoki* (Rangitoki) Fragment: Further Analysis of a Short *rongorongo* Sequence on Bark-cloth from Easter Island." In: *Asian and African Studies (Bratislava)* 29.1, pp. 26–41 and 113–118.

Schomaker, Lambert and Marius Bulacu (2004). "Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Upper-Case Western Script." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.6, pp. 787–798.

Sproat, Richard W. (2000). *A Computational Theory of Writing Systems*. Cambridge: Cambridge University Press.

——— (2003). "Approximate String Matches in the RR Corpus." URL: http://serrano.ai.uiuc.edu/rws/ror/ (visited on 22 October 2007).

——— (2007). "LSA 369: *Rongorongo*." URL: http://catarina.ai.uiuc.edu/LSA07/rongorongo.html (visited on 5 November 2007).

——— (2010). *Language, Technology, and Society*. Oxford and New York: Oxford University Press.

——— (2013). "Corpora and Statistical Analysis of Non-Linguistic Symbol Systems." URL: http://rws.xoba.com/monograph.pdf (visited on 15 February 2014).

——— (2014). "A Statistical Comparison of Written Language and Nonlinguistic Symbol Systems." In: *Language* 90.2, pp. 457–481.

Stimson, Frank J. (1928). "Tahitian Names for the Nights of the Moon." In: *Journal of the Polynesian Society* 37.147, pp. 326–337.

Tacitus, P. Cornelius (ca. 98 CE). "Agricola (De vita et moribus Iulii Agricolae)." in: Book 1. 30. https://www.sacred-texts.com/cla/tac/ag01030.htm. (Visited on 15 May 2019).

Tambiah, Stanley J. (1968). "The Magical Power of Words." In: *MAN (The Royal Anthropological Institute of Great Britain and Ireland)* 3.2, pp. 175–208.

Thomson, William J. (1891). "Te Pito te Henua, or Easter Island." In: *Annual Reports of the Smithsonian Institution for 1889*. Report of the United States National Museum for the Year Ending June 30, 1889. Washington: Smithsonian Institution, United States National Museum, pp. 447–552.

Tilley, Christopher (1991). *Material Culture and Text: The Art of Ambiguity*. London and New York: Routledge.

Tuldava, Juhan (2005). "Stylistics, Author Identification." In: *Quantitative Linguistik / Quantitative Linguistics: Ein internationales Handbuch / An International Handbook*. Ed. by R. Köhler, G. Altmann, and R.G. Piotrowski. Berlin / New York: Walter de Gruyter, pp. 368–387.

van Hoorebeeck, Albert (1979). *La Vérité sur l'île de Pâques*. Le Havre: Pierrette d'Antoine.

van Rijsbergen, Cornelis J. (1979). *Information Retrieval*. 2nd ed. London: Butterworths.

Vogler, Erika (1989). "Die Europäer entdecken die Osterinsel." In: *1500 Jahre Kultur der Osterinsel: Schätze aus dem Land des Hotu Matua*. Ed. by Heide-Margaret Esen-Baur. Mainz: Philipp von Zabern, pp. 53–79.

von Kotzebue, Otto (1825). *Entdeckungsreise in die Südsee und nach der Berings-Straße zur Erforschung einer nordöstlichen Durchfahrt. Unternommen in den Jahren 1815, 1816, 1817 und 1818 auf Kosten Sr. Erlaucht des Herrn Reichskanzlers Grafen Rumanzoff auf dem Schiffe Rurick unter dem Befehle des Lieutenants der Russisch-kaiserlichen Marine*. Wien: Kaulfuß und Krammer.

Wallin, Paul and Helene Martinsson-Wallin (2001). "The 'Fish' for the Gods." In: *Rapa Nui Journal* 15.1, pp. 7–10.

Wang, James Z. et al. (2010). "Determining the Sexual Identities of Prehistoric Cave Artists Using Digitized Handprints: A Machine Learning Approach." In: *Proceedings of the 18th International Conference on Multimedia, Firenze, Italy, October 25–29, 2010*.

Wang, Tao (2007). "Ancient Writing: China's 'Dead Sea Scrolls'." In: *Discovery! Unearthing the New Treasures of Archaeology*. Ed. by Brian Fagan. London: Thames & Hudson, pp. 240–241.

Waterfield, Hermione (2006). "Harry Geoffrey Beasley, 18 December 1881 to 24 February 1939." In: *Provenance: Twelve Collectors of Ethnographic Art in England, 1760–1990*. Ed. by Hermione Waterfield and J. C. H. King. Geneva: Somogy Art Publishers, pp. 78–91.

Wettler, Manfred, Reinhard Rapp, and Peter Sedlmeier (2005). "Free Word Associations Correspond to Contiguities between Words in Texts." In: *Journal of Quantitative Linguistics* 12.2, pp. 111–122.

Wieczorek, Rafał M. (2013). "Naming the *rongorongo* Artifacts." Manuscript in the collection of TSM.

————— (2016a). "*Rongorongo* tablet Keiti: Does it Contain Astronomical Instructions?" In: *Rapa Nui—Easter Island: Cultural and Historical Perspectives*. Ed. by Ian Conrich and Hermann Mückler. Berlin: Frank & Timme, pp. 183–200.

————— (2016b). "Two Unusual *moko* Figurines from the Peabody Essex Museum in Salem." In: *Rapa Nui Journal* 30.1, pp. 13–18.

Wieczorek, Rafał M. and Paul Horley (2015). "The Replicas of *rongorongo* Objects in the Musée du Quai Branly (Paris)." In: *Journal de la Société des Océanistes* 140.1, pp. 123–142.

Williams, H.W. (1928). "The Nights of the Moon." In: *The Journal of Polynesian Society* 37.147, pp. 338–356.

Winn, Shan M. (1981). *Pre-Writing in Southeastern Europe: The Sign System of the Vinča Culture, ca. 4000 BC*. Calgary: Western Publishers.

Quantifying Sound-Graphic Systematicity

Application to Multiple Phonographic Orthographies

Hana Jee · Monica Tamariz · Richard Shillcock

Abstract. Do letter-shapes predict in any way the canonical sounds they represent? Does the letter <a> in any sense visually predict its canonical pronunciation /æ/? We extended existing quantitative approaches to measuring systematicity between phonology and semantics. We quantified all pairwise visual distances between letters, using Hausdorff distance. We took the corresponding canonical pronunciations of the letters and quantified all pairwise distances between their feature-level representations, using edit distance and Euclidean distance. We defined letter-sound systematicity as a correlation between these two lists of distances. We confirmed Korean as the gold standard for letter-sound systematicity; it was designed in the 15th century to have exactly this characteristic. We found small but significant correlations in Arabic, Cyrillic, English, Finnish, Greek and Hebrew orthographies, with Courier New giving the most consistent correlations. Pitman's English shorthand and the Shavian alphabet also showed robust systematicity, and baseline fictitious orthographies showed no systematicity, validating our approach.

1. Background

It is a natural question whether certain parts of a letter or character are topologically related to its meaning or sound. This idea was in fact realized as hieroglyphs or logographs whose written characters are visually

Hana Jee ⬩ 0000-0001-6248-9786

Psychology, School of Philosophy Psychology and Language Sciences, The University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, Scotland

E-mail: hana.jee@ed.ac.uk

Monica Tamariz ⬩ 0000-0003-4688-1774

Psychology, School of Philosophy Psychology and Language Sciences, Heriot-Watt University

E-mail: m.tamariz@hw.ac.uk

Richard Shillcock ⬩ 0000-0002-0616-3703

Psychology, School of Philosophy Psychology and Language Sciences, The University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, Scotland

E-mail: rcs@inf.ed.ac.uk

iconic. For example, the Chinese character <人> 'man' changes its size and location—as in <从> 'to follow' or in <囚> 'to lock up'—maintaining its original meaning in the different contexts. This iconicity facilitates learning the orthography (Dingemanse et al., 2015). For example, the location of the dot distinguishes the meanings between <犬> 'a dog' and <太> 'huge'. It is hard to explain why it is not the other way around, until one knows the former character visually represents a dog wagging its tail.

Phonographs allow far more room for arbitrariness between letters and the corresponding sound unit, but even phonograph users have attempted to theorize about letter shapes in a similar manner: the Roman letter <A> represents a bull's horn upside down; <O> represents the mouth shape of /o:/; and as <S> looks like a snake, it naturally sounds /s/ (Robinson, 1995). The reason why these speculations remained as speculations is related to the problem of this sort of rationalization: there is no consistent theory to apply to all the letter shapes. Such explanations seem to be based on somewhat haphazard analogy.

What is the origin of this propensity to think that meaning inheres in unmotivated written symbols? Looking at how writing emerged may provide an answer. Visual representation started with describing concrete objects (ibid.), but this must have involved some larger semantic value than the object itself—an intention, for example. It is likely that the anonymous painter of the Great Black Bull in the Lascaux cave retrieved the impression of a bull when wishing for a successful hunt.

There are a few scenarios regarding the emergence of writing. One of them suggests that the emergence of agriculture required record-keeping. Instead of a hand-to-mouth lifestyle, people had to remember, for example, the amount of a harvest and the proportion of seed-corn (Schmandt-Besserat, 1989). Many researchers agree that the act of writing began for business purposes: as communities grew and cities were formed, larger scale trade appeared (Havelock, 1976; Robinson, 1995; Rogers, 2005). These proto-writings (Robinson, 1995) occurred in various media, like sticks with notches, clay tokens, and numerical tablets, implying the necessity of simpler, quicker recording. At the same time, administrative procedures, such as those involving tax and the distribution of the population, were required, as in the Sumerian capital, Uruk (Sampson, 1985). These all indicate that the first writing involved forms of numbers—abstract concepts but unmotivated logographs (Havelock, 1976; Robinson, 1995; Rogers, 2005; Schmandt-Besserat, 1989). The ruins of the Assyrian empire (the first millennium BC) showed that their writing did not resemble pictography any more (Robinson, 1995). They managed to establish an arbitrary connection between written symbols and their connotations. However, these symbols did not yet acquire the status of phonographs, not being connected to individual sound units.

The idea of phonographic symbols appeared only after the discovery of the 'Rebus principle' (ibid.; Rogers, 2005), where a single pictographic symbol could be connected to a sound value. The sound unit at this stage was not necessarily phonemic and more likely syllabic (Havelock, 1976; Robinson, 1995). The more specific, phonemic association required the ability to segment the continuous flow of vocal sounds and to realize the differences between air flows and articulatory obstructions (Havelock, 1976). The presence or absence of vibration in the larynx (voiced vs. voiceless) also had to be noticed.

Once an alphabet set is established, the constituents of the system need to be in balance between *efficiency* and *distinctiveness*: they should be easy to write and distinguished from one another. For extreme efficiency, all the letters might look the same but at the cost of distinction. At the other extreme, the letters might differ in shape size, colour and material as well as orientation. In fact, many phonographic orthographies in human history satisfy *coherent discrimination* among the constituents of the system by addition, subtraction, duplication, or orientation change. It is plausible to expect these scripts to have developed some sort of *systematicity*, in order to make the best use of the limited resources to facilitate acquisition and transmission of the orthography.

As an exclusively cultural heritage (Sampson, 1985), each writing script undergoes its own cultural evolution. Removing inefficiencies and not creating a new revolutionary feature, is the central role of cultural evolution (ibid.), a process that can be enhanced by repetitions and extensive communication. Multiple factors condition letter shapes. The nature of writing materials (Sirat, 1994) decides the angularity of letters: C versus <. The combination of writing materials and writing postures also affect the complexity of letters: compare a pen on a paper and a chisel on a clay tablet. Watt (2013) pointed out that letters tend to face the same direction; the facing direction is defined as the direction of ornaments and headings—for example, Arabic numbers mostly face left. Anecdotal evidence says that children often reverse, for instance, the letter B until they subconsciously understand that asymmetric English letters generally face rightwards. The direction of the script can be also changed for political and cultural reasons. People in conquered territories often had to adapt to a new writing custom. For example, Egyptians began to write from left to right when they accepted Christianity but later returned to write from right to left when Islam prevailed in the region in the 7C. The direction of script affects the direction of letters because moving backwards slows the pace of continuous writing, reducing efficiency. Hebrew seems to take longer to write because the overall script moves from right to left whereas horizontal strokes are written from left to right (Sirat, 1994). In Rome and Greece, the scripts were written successively from right to left, and then left to right, termed boustrophedon fashion or ox-turning. Asymmetric letters like B, E, N

were frequently written in their mirror images to match the direction of the script. Some of the letters of the modern Roman alphabet therefore remained as their mirror images when the writing direction was stabilized from left to right (Sirat, 1994).

Watt (1979) introduced four hypothetical forces that affect letter shapes. *Homogenization* means the letters look more alike; *heterogenization* means they are distinguished from each other. For example, the uppercase letters <D>, <E>, <F>, <H> are visually homogeneous whereas their lowercase counterparts <d>, <e>, <f>, <h> are considerably heterogenized. *Facilitation* means the tendency for letters to be easy to produce. For instance, cursive movement minimizes direction shifts and hand movements, for greater writing speed (Sirat, 1994). Finally, *inertia*, is a conservative force to stabilize the system. These forces are more topological than kinetic. When they are in equilibrium, the orthography system stays the same, but when any of the first three forces gets stronger, letter shapes may change. Any change of letter shapes or introduction of a new letter occurs *in connection with* the other elements in the system, the other letter shapes and sounds (Brekle, 1994; Watt, 1979, Watt, 1994, Watt, 2013).

In this paper, we suggest a novel approach to investigating the assumed systematicity between letters and sounds. It is, however, not a symbolic logic in which "letter <g> has a definite feature of its sound /g/". It is rather a reflection of the system as a whole: "is letter <g> close to letter <k> as much as the sound /g/ is close to the sound /k/? How much do the distances among the phonemes correlate with the distances among the visual representations of those phonemes? To our knowledge, this is the first such quantitative demonstration of letter-sound systematicity across the whole alphabet.

We transferred this method from recent studies reporting systematicity between semantics and phonology (Dautriche, Mahowald, Gibson, and Piantadosi, 2017; Monaghan, Shillcock, Christiansen, and Kirby, 2014; Shillcock, Kirby, McDonald, and Brew, 2001; Tamariz, 2008). We explored the systematic relation between phonology and orthography in Arabic, Cyrillic, English, Finnish, Greek, Hebrew, and Korean.

2. Procedure

We measured all the pairwise visual distances between letters and the corresponding pairwise phonological distances between the canonical pronunciations of those letters, in the respective alphabets. The total pairwise distances in phonology or semantics are $N \times (N - 1)/2$. We defined letter-sound systematicity as the correlation between the resulting two lists of distances, as in the Mantel Test (Mantel, 1967). The significance of the correlation between these two lists of pairwise distances

was tested with a Monte-Carlo permutation test, as in the published literature on word-level systematicity. The whole process was conducted in Python 3.7.1.[1]

2.1.   Phonological Distances

We encoded the phonemes of each language into feature vectors (cf. Farmer, Christiansen, and Monaghan, 2006) based on the International Phonetic Alphabet (IPA). The features consisted of place and manner of articulation. We marked 1 if a phoneme had the feature and 0 if it did not, and transformed each phoneme into a binary vector. For example, /b/ can be represented as [0,1,0,0,1,0,0]: palatal, labial, dental, throat, plosive, affricate, and fricative. The length of the vectors equalled the total number of phonological features of a language.

We measured the distances between two vectors as feature edit distance, which counts the number of features different between the two vectors, and as Euclidean distance, which measures the shortest geometric distance between two vectors. (Multiple distance metrics demonstrate the robustness of the results.) The more dissimilar two vectors are, the larger the values that are returned. For all phonological distance measures, we used `textdistance` 4.1.4 (Python 3.7.1)[2].

2.2.   Orthographical Distances

We measured the distances between two letter images by Hausdorff distance (Huttenlocher, Klanderman, and Rucklidge, 1993). Hausdorff distance measures the difference between two images by first comparing each pixel of 'X' and 'Y' and then calculating Euclidean distance between the pixel from 'X' and the closest pixel from 'Y'. Being fundamentally asymmetric— the distance from 'X' to 'Y' is different from 'Y' to 'X'— the larger value is used by definition.

Because the letters were treated as images, different fonts returned different results. We examined various fonts available in Microsoft including serif, sans-serif, and cursive fonts: 29 fonts for Cyrillic, English, Finnish, and Greek (Table 1); 10 fonts for Arabic (Table 2); 13 fonts for Hebrew (Table 3); and 88 fonts for Korean (Appendix F). The letters were all centrally aligned with the default font setting and saved as an identically sized PNG image file. An implementation of Hausdorff distance in Python 3.7.1 converted these images into black and white raster graphics and returned numeric values as results.

---

1. The Python code are available from `https://github.com/HanaJee/hausdorff-distance-letters.git`.

2. Textdistance 4.1.4 imported from `https://pypi.org/project/textdistance/` in July 2019.

TABLE 1. Fonts examined for Cyrillic, English, Finnish, and Greek

Serif fonts	Book Antiqua, Cambria, Constantia, Courier New, *Gabriola,* Georgia, Lucida Console, Palatino Linotype, Times New Roman
Sans-serif fonts	Arial, **Arial Black,** Candara, Calibri, Calibri Light, Century Gothic, *Comic Sans MS*, Consolas, Corbel, **Franklin Gothic Medium**, **Impact**, Lucida Sans Unicode, Microsoft Sans Serif, Segoe UI Symbol, Tahoma, Trebuchet MS, Verdana
Cursive style	*Lucida Handwriting*, **Segoe Print,** **Segoe Script**

2.3. Samples

Arabic

A written Arabic alphabet (Arabic abjad) can have a maximum of four different forms: in the initial positions, in the middle of a word, in the final positions, and in the isolated forms (Erfani, 2005). We examined the isolated forms as they are the canonical letters that are first taught to children. Note that Arabic long vowels (<ا> /aː/, <و> /w/, and <ي> /j/) are included in the set of the alphabet, whereas short vowels (<◌ُ> /u/, <◌َ> /a/, and <◌ِ> /i/) are considered diacritics. We collected 28 Arabic letters and vectorized their corresponding phonemes based on 18 IPA features (Appendix A).

Cyrillic

Cyrillic script is used in many Eastern European countries, including Russia, but there are variations. Russian Cyrillic, for example, was reformed in the 18th century. Ukrainian, Bulgarian, Serbian, Macedonian, and Iranian script among others also look slightly different. We used the common Cyrillic letters and their phonemes (Appendix B). The letters <Е>, <Ю> and <Я> were excluded because they are diphthongs (/jɛ/, /ju/ and /ja/, respectively), as was <Ь>, because it simply makes consonants softer and does not have any phonetic value. Accordingly, we made 25 phoneme vectors based on 20 IPA features (Appendix B).

English

As a deep orthography (Seymour, Aro, and Erskine, 2003), English letters are linked to more than one phoneme. We first constrained the

sound of a letter according to the British phonics approach (Lloyd, Wernham, Jolly, and Stephen, 1998). Phonics teaches children the most frequent and canonical sound of the letter. We excluded <x> and <q> from the sample because the former is a polyphone /ks/ and the latter almost always co-occurs with <u>. In total, 24 letters were converted into feature-vectors, taken from Harm and Seidenberg (1999).

Finnish

Finnish script is the same as English except for three additional letters: <ä>, <ö>, and <å>, and the letters <k>, <p>, and <t> have tensed sounds, not aspirated. We included <q> because it is independently pronounced /k/. The 28 letters and 17 phonetic features are listed in Appendix C.

Greek

Greek uppercase letters are historically important in that they are closely related to ancient orthographies such as Phoenician. Lowercase letters have distinct forms from uppercase letters (Appendix D). The uppercase letter <Σ> (/s/) corresponds to two lowercase letters, which we included. We excluded <Ξ> and <Ψ>, as well as their corresponding lowercases <ξ> and <ψ>, because they are diphthongs: /ks/ and /ps/, respectively. We used 19 IPA features for the Greek phonemes.

Hebrew

As a consonantal orthography, written Hebrew for advanced readers does not indicate vowel values. The vowels are only written out for children and foreign learners until they get used to reading. We examined 33 consonants with 14 IPA phonetic features (Appendix E).

Korean

Hangeul, the Korean orthography, was artificially invented in the 15th century. It is well known for its one-to-one connection between letters and sounds, and for the fact that its letters were designed based on the shape of articulation. For example, <ㄱ> /g/ represents the tongue touching the soft palate. Korean phonology distinguishes between phonemes that are considered allophones by English speaker: /p/ in 'pie' and 'spy' are perceived as *aspirated* and *tensed*, respectively. Along with the *lenis* sound that shares the same articulation point without aspiration, these phonemes have visually systematic forms (e.g., <ㅂ> /b/ -<ㅍ> /p/ -<ㅃ> /p*/). Based on more cultural grounds, Korean written

vowels are composed of three components: <·>, <⌐>, and < ⌐ >, which respectively represent the heaven, earth and human. In total, we examined 16 consonants and 10 monophthongs (Appendix F).

Other Orthographies

We additionally examined four ancient Semitic orthographies (Phoenician, Nabataean, Early Arabic, and Aramaic), two English substitute systems (*Pitman's shorthand* and the *Shavian alphabet*) and two fictitious orthographies (*Aurebesh* from *Star Wars* and *Klingon* from *Star Trek*) in terms of sound-letter systematicity. We expect if such a correlation is found in the modern conventional orthographies, it evolved over cultural time. We do not expect to observe any sound-letter systematicity in the fictitious systems that have not undergone natural selection in human culture. Finally, the artificially, consciously constructed letters in the Pitman's shorthand and Shavian alphabet may be expected to have a systematicity comparable to Korean orthography.

3. Results

General Results

For each orthography, we calculated systematicity as Pearson's r and confirmed the significance level with Monte-Carlo permutation tests. For each of the naturally occurring orthographies there were fonts for which significant systematicity obtained: for Korean 85 out of 88 fonts produced significant systematicity; for Finnish only 2 fonts out of 29 returned a significant systematicity. When a font exhibited significant systematicity, it was generally of the order of $r = 0.1 - 0.15$ (see Fig. 1; see below, also); similar letters tend to have similar sounds. Greek lower cases, in contrast, showed a negative correlation; similar letters tend to have distinct sounds.

Arabic

Table 2 indicates that Arabic letters tend to correlate with their sounds. Simplified Arabic consistently showed significant systematicity regardless of phonemic distance measure.

Cyrillic

Cyrillic upper and lower cases both correlated with the phonemes only in Courier New. The upper cases: $r = .14$, $p = .02$ when measured by Euclidean distance, $r = .18$, $p < .01$ when measured by feature edit distance. The lower cases: $r = .14$, $p = .02$ when measured by Euclidean distance and $r = .18$, $p < .001$ when measured by feature edit distance.
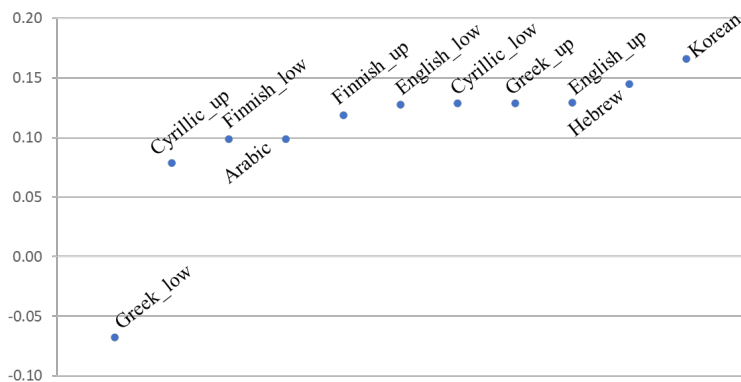
FIGURE 1.   Letter-sound correlations of the conventional orthographies:
we averaged the correlation coefficients from various fonts only when p-value < .1.

TABLE 2.   The letter-sound correlations in 10 Arabic fonts. Note: * $p < .05$, **
$p < .01$, *** $p < .001$, $N = 378$, phonological distance $M = 1.46$, $SD = 0.27$ (Eu-
clidean); $M = 2.46$, $SD = 1.43$ (feature edit); orthographical distance $M = 10.50$,
$SD = 2.64$

Font	Example	Euclidean distance			feature edit distance		
		r	p-value		r	p-value	
Simplified Arabic	تشرفت بمقابلتك	0.15	< .001	***	0.1	0.05	*
Arial Black	تشرفت بمقابلتك	0.13	0.01	**	0.09	0.07	
Times New Roman	تشرفت بمقابلتك	0.13	0.01	**	0.09	0.07	
Arabic Typesetting	تشرفت بمقابلتك	0.12	0.02	*	0.02	0.74	
Traditional Arabic	تشرفت بمقابلتك	0.11	0.03	*	0.04	0.42	
Courier New	تشرفت بمقابلتك	0.07	0.2		0.05	0.32	
Microsoft Sans Serif	تشرفت بمقابلتك	0.05	0.37		0.12	0.02	**
Segoe UI	تشرفت بمقابلتك	0.04	0.38		0.12	0.02	**
Andalus	تشرفت بمقابلتك	0.04	0.45		0.05	0.29	
Tahoma	تشرفت بمقابلتك	-0.01	0.83		0.11	0.03	**

English

For upper cases, Cambria consistently returned correlation: $r = .11$,
$p = .07$ when measured by Euclidean distance, $r = .12$, $p = .04$ when
measured by feature edit distance. Gabriola ($r = .12$, $p = .04$), Georgia
($r = .10$, $p = .08$), and Impact ($r = .15$, $p = .01$) additionally showed the
result when measured by feature edit distance only. For lower cases,
Franklin Gothic Medium ($r = .15$, $p = .01$), Arial Black ($r = .14$, $p = .02$),

Verdana ($r = .14$, $p = .02$), Cambria ($r = .13$, $p = .03$), and Tahoma ($r = .12$, $p = .04$) returned the results only when measured by feature edit distance.

Finnish

Significant systematicity was found only in uppercase Courier New ($r = .12$, $p = .02$). For lower cases, Segoe Script ($r = .10$, $p = .05$ measured by Euclidean distance) and Trebuchet MS ($r = .11$, $p = .04$, measured by feature edit distance) returned coefficient above significance level.

Greek

For upper cases, Courier New consistently showed robust coefficient: $r = .15$, $p = .02$ measured by Euclidean distance, $r = .16$, $p = .02$ measured by feature edit distance. Book Antiqua ($r = .13$, $p = .04$) was also significant when measured by Euclidean distance. Although marginal, the lowercase Courier New returned the negative correlation ($r = -.11$, $p = .09$) when measured by Euclidean distance.

Hebrew

All 13 fonts returned highly significant correlation coefficients (Table 3). Some fonts returned letter-sound systematicity even higher than that of Korean orthography.

Korean

Almost all 88 Korean fonts returned significant letter-sound correlation, including a few representative fonts: 굴림: $r = .24$, $p < .001$; 바탕: $r = .18$, $p < .001$; 궁서: $r = .30$, $p < .001$; 맑은고딕: $r = .18$, $p < .001$. KCC 은영 returned the highest coefficient: $r = .39$, $p < .001$. We re-calculated the correlation excluding each letter to investigate which contributes the most to the whole correlation. Each letter seems to contribute approximately equally to the whole letter-sound correlation.

Other Orthographies

None of the four ancient orthographies returned significant systematicity; nor did the two fictitious orthographies. We conducted Monte-Carlo permutation tests for verification.

The two English substitute writing systems returned high positive letter-sound correlations: *Pitman's shorthand*, in which $r = .35$, $p < .001$; and the *Shavian alphabet*, $r = .2$, $p < .001$.

TABLE 3. The letter-sound correlation r in 13 Hebrew fonts (all $p-value < .001$). Note: * $p < .05$, ** $p < .01$, *** $p < .001$, $N = 1035$, phonological distance $M = 1.91$, $SD = 0.43$ (Euclidean); $M = 3.14$, $SD = 1.79$ (feature edit); orthographical distance $M = 15.57$, $SD = 9.89$

Font	Example	Euclidean distance	Feature edit distance
Levenim MT	שלום	0.35	0.31
Narkisim	שלום	0.35	0.31
Miriam	שלום	0.34	0.31
Times New Roman	שלום	0.34	0.31
David	שלום	0.33	0.3
Lucida Sans Unicode	שלום	0.30	0.28
Gisha	שלום	0.27	0.27
Arial	שלום	0.25	0.26
Arial Black	שלום	0.25	0.26
Calibri Light	שלום	0.25	0.26
Microsoft Sans Serif	שלום	0.24	0.27
Courier New	שלום	0.21	0.23
Tahoma	שלום	0.18	0.24

4.   Discussion

We explored the systematicity of letter-sound mapping over a number of orthographies, from a new perspective. Seven conventional orthographies, as well as two reformed spelling systems, demonstrated that letters to some extent correlate with their pronunciations. *Hangeul*, the systematically invented orthography with a sophisticated understanding of phonology, constitutes the highest benchmark of letter-sound correlation; other artificial orthographies, *Pitman's shorthand* and the *Shavian alphabet* returned a similarly high correlation. Letter-sound correlation increases when visually similar figures are linked to articulatorily similar phonemes (e.g., <כ> /k/ - <ג> /g/ or <ת> /t/ - <ד> /d/). This fact explains why Hebrew also demonstrated a high correlation. The visual difference of letter shapes efficiently categorise the place of articulation and distinguish voiceless from voiced sounds (e.g., /k/ - /x/ or /v/ - /b/ in Appendix 5). The systematicity of an orthography is enhanced when adding or subtracting a stroke or the orientation change of letter shapes occurs systematically with the corresponding phoneme pairs (e.g., voiced-voiceless).

Comparatively low coefficients of the orthographies with Roman alphabets (Cyrillic, English, Finnish, and Greek) may be attributable to their complicated history. They originated from Phoenician alphabets

(1000 BC), known as the first stable alphabetic script (Havelock, 1976; Robinson, 1995). It diverged to Hebrew and Greek, and the latter was borrowed by the Romans. The Roman alphabets spread through Europe and one of the lineages settled down as the English alphabets (Havelock, 1976; Robinson, 1995; Rogers, 2005). Some 3000 years of the history of this *Northwest Semitic Graeco-Roman-Etruscan alphabet* (Havelock, 1976) naturally allowed cultural intervention, sometimes organized (ibid.; Robinson, 1995; Rogers, 2005). For example, when Phoenician 22-consonant alphabets were accepted by Greeks, some phonetic values (mostly weak consonants) were changed to vowels. At the same time, three more vowels were added, resulting in 25 characters in total. Later, Runes, the Germanic alphabets entered Roman culture, influencing some of their letters: r, i, and b. Middle English went through the *Great English Vowel Shift*, as well as the distinction of upper cases from lower cases.

We expected the modern European alphabet systems to demonstrate stronger systematicity than the ancient orthographies. Four ancient orthographies did not show significant systematicity. The authenticity of the phonemes (and characters) recovered (Havelock, 1976; Robinson, 1995) may be not perfectly reliable.

In conclusion, the human brain is adept at taking advantage of any type of systematicity, from the level of the neural substrate to cross-modality processing (Bavelier and Neville, 2002; Spence, 2011). There are many demonstrations of audio-visual multisensory perception (Baier, Kleinschmidt, and Müller, 2006; Calvert, Brammer, et al., 1999; Calvert, Campbell, and Brammer, 2000; Calvert, Hansen, Iversen, and Brammer, 2001; Fiebelkorn, Foxe, and Molholm, 2010; Kriegstein and Giraud, 2006; Zangenehpour and Zatorre, 2010), some of which specifically focus on grapheme-phoneme relations (Raij, Uutela, and Hari, 2000; Atteveldt, Formisano, Goebel, and Blomert, 2004; Weissman, Warner, and Woldorff, 2004). Although the data generally imply that no area is exclusively related to reading, the human brain certainly has the wherewithal to take advantage of the type of systematicity we have demonstrated in the relation between letters and their canonical pronunciation. One potential process underlying the emergence of systematicity may be Zipf's principle of least effort (Zipf, 1949), whereby least effort in pronunciation travels with least effort in writing a character, with these processes conditioning the pairwise distances within phonological and visual space.

References

Atteveldt, N. van et al. (2004). "Integration of letters and speech sounds in the human brain." In: *Neuron* 43.2, pp. 271–282.

Baier, B., A. Kleinschmidt, and N.G. Müller (2006). "Cross-modal processing in early visual and auditory cortices depends on expected statistical relationship of multisensory information." In: *Journal of Neuroscience* 26.47, pp. 12260–12265.

Bavelier, D. and H.J. Neville (2002). "Cross-modal plasticity: where and how?" In: *Nature Reviews Neuroscience* 3.6, pp. 443–452.

Brekle, H.E. (1994). "Some thoughts on a historico-genetic theory of the lettershapes of our alphabet." In: *Writing systems and cognition: Perspectives from psychology, physiology, linguistics, and semiotics*. Ed. by W. Watt. Berlin: Springer.

Calvert, G.A., M.J. Brammer, et al. (1999). "Response amplification in sensory-specific cortices during crossmodal binding." In: *Neuroreport* 10.12, pp. 2619–2623.

Calvert, G.A., R. Campbell, and M.J. Brammer (2000). "Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex." In: *Current Biology* 10.11, pp. 649–657.

Calvert, G.A. et al. (2001). "Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect." In: *Neuroimage* 14.2, pp. 427–438.

Dautriche, I., E. Chemla, and A. Christophe (2016). "Word learning: Homophony and the distribution of learning exemplars." In: *Language Learning and Development* 12.3, pp. 231–251.

Dautriche, I. et al. (2017). "Wordform similarity increases with semantic similarity: An analysis of 100 languages." In: *Cognitive science* 41.8, pp. 2149–2169.

Dingemanse, M. et al. (2015). "Arbitrariness, iconicity, and systematicity in language." In: *Trends in Cognitive Sciences* 19.10, pp. 603–615.

Erfani, I.M. (2005). *Arabic for Beginners*. Ottawa: Laurier Books Limited.

Farmer, T.A., M.H. Christiansen, and P. Monaghan (2006). "Phonological typicality influences on-line sentence comprehension." In: *Proceedings of the National Academy of Sciences* 103.32, pp. 12203–12208.

Fiebelkorn, I.C., J.J. Foxe, and S. Molholm (2010). "Dual mechanisms for the cross-sensory spread of attention: how much do learned associations matter?" In: *Cerebral Cortex* 20.1, pp. 109–120.

Harm, M.W. and M.S. Seidenberg (1999). "Phonology, reading acquisition, and dyslexia: insights from connectionist models." In: *Psychological review* 106.3, pp. 491–528.

Havelock, E.A. (1976). *Origins of Western Literacy. Four Lectures delivered at the Ontario Institute for Studies in Education*. Toronto: Ontario Institute for Studies in Education.

Huttenlocher, D.P., G.A. Klanderman, and W.J. Rucklidge (1993). "Comparing images using the Hausdorff distance." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15.9, pp. 850–863.

Kriegstein, K. von and A.L. Giraud (2006). "Implicit multisensory associations influence voice recognition." In: *PLoS biology* 4.10.

Lloyd, S. et al. (1998). *The phonics handbook*. Chigwell: Jolly Learning.

Mantel, N. (1967). *The detection of disease clustering and a generalized regression approach. Cancer Research*.

Monaghan, P. et al. (2014). "How arbitrary is language?" In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1651.

Raij, T., K. Uutela, and R. Hari (2000). "Audiovisual integration of letters in the human brain." In: *Neuron* 28.2, pp. 617–625.

Robinson, A. (1995). *The story of writing*. London: Thames and Hudson.

Rogers, H. (2005). *Writing systems: A linguistic approach*. Oxford: Blackwell.

Sampson, G. (1985). *Writing systems*. London: Hutchinson.

Schmandt-Besserat, D. (1989). *Two precursors of writing: Plain and complex tokens. The origins of writing*.

Seymour, P.H., M. Aro, and J.M. Erskine (2003). "Collaboration with COST Action A8 Network." In: *British Journal of psychology* 94.2, pp. 143–174.

Shillcock, R.C. et al. (2001). "Filled pauses and their status in the mental lexicon." In: *Proc. 2001 Conf. of Disfluency in Spontaneous Speech*, pp. 53–56.

Sirat, C. (1994). "Handwriting and the writing hand." In: *Writing systems and cognition: Perspectives from psychology, physiology, linguistics, and semiotics*. Ed. by W. Watt. Berlin: Springer.

Spence, C. (2011). "Crossmodal correspondences: A tutorial review. Attention, Perception, and." In: *Psychophysics* 73.4, pp. 971–995.

Tamariz, M. (2008). "Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon." In: *The Mental Lexicon* 3.2, pp. 259–278.

Watt, W.C. (1979). "Iconic equilibrium." In: *Semiotica* 28.1–2, pp. 31–62.

——— (1994). *Curves as angles. In Writing systems and cognition: Perspectives from psychology, physiology, linguistics, and semiotics (Vol*. Ed. by W. Watt. Berlin: Springer.

——— (2013). *Writing systems and cognition: Perspectives from psychology, physiology, linguistics, and semiotics (Vol. 6)*. Berlin: Springer.

Weissman, D.H. van, L.M. Warner, and M.G. Woldorff (2004). "The neural mechanisms for minimizing cross-modal distraction." In: *Journal of Neuroscience* 24.48, pp. 10941–10949.

Zangenehpour, S. and R.J. Zatorre (2010). "Crossmodal recruitment of primary visual cortex following brief exposure to bimodal audiovisual stimuli." In: *Neuropsychologia* 48.2, pp. 591–600.

Zipf, G.K. (1949). *Human behavior and the principle of least-effort*. Reading: Addison-Wesley.

A. Arabic

TABLE 4. Arabic letters and their phonemes

letter	phoneme	letter	phoneme	letter	phoneme	letter	phoneme
ا	/a/	د	/d/	ض	/dˤ/	ك	/k/
ب	/b/	ذ	/ð/	ط	/tˤ/	ل	/l/
ت	/t/	ر	/r/	ظ	/ðˤ/	م	/m/
ث	/θ/	ز	/z/	ع	/ʕ/	ن	/n/
ج	/dʒ/	س	/s/	غ	/ɣ/	ه	/h/
ح	/ħ/	ش	/ʃ/	ف	/f/	و	/w/
خ	/x/	ص	/sˤ/	ق	/q/	ي	/j/

TABLE 5. The features of Arabic phonemes

Letters	Phonemes	Place of Articulation									Manner of Articulation								Vowel quality
		Labial	Dental	Denti-alveolar Plain	Denti-alveolar Emphatic	Palatal	Velar	Uvular	Pharyngeal	Glottal	Nasal	Stop Voiceless	Stop Voiced	Fricative Voiceless	Fricative Voiced	Affricate Voiced	Trill	Approximant	
ا	a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ب	b	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ت	t	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ث	θ	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ج	dʒ	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
ح	ħ	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
خ	x	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
د	d	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ذ	ð	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ر	r	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ز	z	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
س	s	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ش	ʃ	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
ص	sˤ	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ض	dˤ	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
ط	tˤ	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ظ	ðˤ	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ع	ʕ	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
غ	ɣ	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
ف	f	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ق	q	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
ك	k	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
ل	l	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
م	m	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
ن	n	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
ه	h	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
و	w	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
ي	j	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0

B.   Cyrillic

Table 6.   Cyrillic letters and their phonemes

Upper	Lower	Phoneme	Upper	Lower	Phoneme
А	а	/a/	О	о	/o/
Б	б	/b/	П	п	/p/
В	в	/v/	Р	р	/r/
Г	г	/ɡ/	С	с	/s/
Д	д	/d/	Т	т	/t/
Ж	ж	/ʒ/	У	у	/u/
З	з	/z/	Ф	ф	/f/
И	и	/i/	Х	х	/x/
Й	й	/j/	Ц	ц	/ts/
К	к	/k/	Ч	ч	/tʃ/
Л	л	/l/	Ш	ш	/ʃ/
М	м	/m/	Щ	щ	/ʃt/
Н	н	/n/			

Table 7.   The features of Cyrillic phonemes

Upper letters	Lower letters	Phonemes	Place of Articulation							Manner of Articulation						Vowel Qualities						
			Voiced	Bilabial	Labio-dental	Alveolar	Palatal	Velar	Post-alveolar	Nasal	Plosive	Affricate	Fricative	Trill	Approximant	Close	Mid	Open	Front	Central	Back	Roundness
А	а	a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
Б	б	b	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
В	в	v	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Г	г	ɡ	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Д	д	d	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Ж	ж	ʒ	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
З	з	z	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
И	и	i	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
Й	й	j	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
К	к	k	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Л	л	l	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
М	м	m	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Н	н	n	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
О	о	o	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1
П	п	p	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Р	р	r	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
С	с	s	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Т	т	t	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
У	у	u	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1
Ф	ф	f	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Х	х	x	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Ц	ц	ts	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Ч	ч	tʃ	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
Ш	ш	ʃ	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
Щ	щ	ʃt	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

C. Finnish

TABLE 8. Finnish letters and their phonemic features

Letter	Phoneme	Place of Articulation						Manner of Articulation					Vowel Quality					
		Voiced	Labial	Alveolar	Palatal	Velar	Glottal	Nasal	Plosive	Fricative	Approximant	Trill	Close	Mid	Open	Front	Back	Roundness
a	ɑ	O	O	O	O	O	O	O	O	O	O	O	O	O	I	O	I	O
b	b	I	I	O	O	O	O	O	I	O	O	O	O	O	O	O	O	O
c	s	O	O	I	O	O	O	O	O	I	O	O	O	O	O	O	O	O
d	d	I	O	I	O	O	O	O	I	O	O	O	O	O	O	O	O	O
e	e	O	O	O	O	O	O	O	O	O	O	O	O	I	O	I	O	O
f	f	O	I	O	O	O	O	O	O	I	O	O	O	O	O	O	O	O
g	g	I	O	O	O	I	O	O	I	O	O	O	O	O	O	O	O	O
h	h	O	O	O	O	O	I	O	O	I	O	O	O	O	O	O	O	O
i	i	O	O	O	O	O	O	O	O	O	O	O	I	O	O	I	O	O
j	j	I	O	O	I	O	O	O	O	O	I	O	O	O	O	O	O	O
k	k	O	O	O	O	I	O	O	I	O	O	O	O	O	O	O	O	O
l	l	O	O	I	O	O	O	O	O	O	I	O	O	O	O	O	O	O
m	m	O	I	O	O	O	O	I	O	O	O	O	O	O	O	O	O	O
n	n	O	O	I	O	O	O	I	O	O	O	O	O	O	O	O	O	O
o	o	O	O	O	O	O	O	O	O	O	O	O	O	I	O	O	I	I
p	p	O	I	O	O	O	O	O	I	O	O	O	O	O	O	O	O	O
q	k	O	O	O	O	I	O	O	I	O	O	O	O	O	O	O	O	O
r	r	O	O	I	O	O	O	O	O	O	O	I	O	O	O	O	O	O
s	s	O	O	I	O	O	O	O	O	I	O	O	O	O	O	O	O	O
t	t	O	O	I	O	O	O	O	I	O	O	O	O	O	O	O	O	O
u	u	O	O	O	O	O	O	O	O	O	O	O	O	I	O	O	I	O
v	ʋ	I	I	O	O	O	O	O	O	O	I	O	O	O	O	O	O	O
w	ʋ	I	I	O	O	O	O	O	O	O	I	O	O	O	O	O	O	O
y	y	O	O	O	O	O	O	O	O	O	O	O	I	O	O	I	O	I
z	z	I	O	I	O	O	O	O	O	I	O	O	O	O	O	O	O	O
ä	æ	O	O	O	O	O	O	O	O	O	O	O	O	O	I	I	O	O
ö	ø	O	O	O	O	O	O	O	O	O	O	O	O	I	O	I	O	I
å	o	O	O	O	O	O	O	O	O	O	O	O	O	I	O	O	I	I

D.   Greek

TABLE 9.   Greek letters and their phonemes

Upper	Lower	Phoneme	Upper	Lower	Phoneme
A	α	/a/	N	ν	/n/
B	β	/v/	O	o	/o/
Γ	γ	/ɣ/	Π	π	/p/
Δ	δ	/ð/	P	ρ	/r/
E	ε	/e/	Σ	σ	/s/
Z	ζ	/z/	Σ	ς	/s/
H	η	/i/	T	τ	/t/
Θ	θ	/θ/	Υ	υ	/i/
I	ι	/i/	Φ	φ	/f/
K	κ	/k/	X	χ	/x/
Λ	λ	/l/	Ω	ω	/o/
M	μ	/m/			

TABLE 10.   The features of Greek phonemes

Upper letter	Lower letter	Phonemes	Place of Articulation							Manner of Articulation				Vowel Qualities							
			Voiced	Bilabial	Labio-dental	Alveolar	Velar	Dental	Nasal	Plosive	Fricative	Lateral approximant	Trill	Close	Close-mid	Mid-back	Open	Central	Front	Back	Roundness
A	α	a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0
B	β	v	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Γ	γ	ɣ	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Δ	δ	ð	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
E	ε	e	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0
Z	ζ	z	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
H	η	i	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
Θ	θ	θ	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
I	ι	i	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
K	κ	k	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Λ	λ	l	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
M	μ	m	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
N	ν	n	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
O	o	o	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
Π	π	p	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
P	ρ	r	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Σ	σ	s	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Σ	ς	s	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
T	τ	t	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Υ	υ	i	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
Φ	φ	f	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
X	χ	x	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Ω	ω	o	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1

E.   Hebrew

TABLE 11. Hebrew letters and their phonemes

Letter	Phoneme	Letter	Phoneme	Letter	Phoneme
א	empty	כ	/k/	פ	/p/
ב	/v/	כ	/x/	פ	/f/
בּ	/b/	ך	/k/	ף	/f/
ג	/g/	ך	/x/	צ	/ts/
ד	/d/	ל	/l/	ץ	/ts/
ה	/h/	מ	/m/	ק	/k/
ו	/v/	ם	/m/	ר	/r/
ז	/z/	נ	/n/	שׁ	/sh/
ח	/x/	ן	/n/	שׂ	/s/
ט	/t/	ס	/s/	תּ	/t/
י	/j/	ע	emp	ת	/t/

TABLE 12. The features of Hebrew phonemes

Letter	Phoneme	Place of Articulation								Manner of Articulation					
		Voiced	Bilabial	Labio-dental	Alveolar	Palatal	Velar	Glottal	Post-alveolar	Nasal	Plosive	Affricate	Fricative	Lateral approx.	Approximant
א	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ב	b	1	1	0	0	0	0	0	0	0	1	0	0	0	0
ב	v	1	0	1	0	0	0	0	0	0	0	0	1	0	0
ג	g	1	0	0	0	0	1	0	0	0	1	0	0	0	0
ד	d	1	0	0	1	0	0	0	0	0	1	0	0	0	0
ה	h	0	0	0	0	0	0	1	0	0	0	0	1	0	0
ו	v	1	0	1	0	0	0	0	0	0	0	0	1	0	0
ז	z	1	0	0	1	0	0	0	0	0	0	0	1	0	0
ח	x	0	0	0	0	0	1	0	0	0	0	0	1	0	0
ט	t	0	0	0	1	0	0	0	0	0	1	0	0	0	0
י	j	1	0	0	0	1	0	0	0	0	0	0	0	0	1
כ	k	0	0	0	0	0	1	0	0	0	1	0	0	0	0
כ	x	0	0	0	0	0	1	0	0	0	0	0	1	0	0
ך	k	0	0	0	0	0	1	0	0	0	1	0	0	0	0
ך	x	0	0	0	0	0	1	0	0	0	0	0	1	0	0
ל	l	0	0	0	1	0	0	0	0	0	0	0	0	1	0
מ	m	0	1	0	0	0	0	0	0	1	0	0	0	0	0
ם	m	0	1	0	0	0	0	0	0	1	0	0	0	0	0
נ	n	0	0	0	1	0	0	0	0	1	0	0	0	0	0
ן	n	0	0	0	1	0	0	0	0	1	0	0	0	0	0
ס	s	0	0	0	1	0	0	0	0	0	0	0	1	0	0
ע	-	0	0	0	0	0	0	0	0	0	0	0	0	0	0
פ	p	0	1	0	0	0	0	0	0	0	1	0	0	0	0
פ	f	0	0	1	0	0	0	0	0	0	0	0	1	0	0
ף	f	0	0	1	0	0	0	0	0	0	0	0	1	0	0
צ	ts	0	0	0	1	0	0	0	0	0	0	1	0	0	0
ץ	ts	0	0	0	1	0	0	0	0	0	0	1	0	0	0
ק	k	0	0	0	0	0	1	0	0	0	1	0	0	0	0
ר	ɣ	1	0	0	0	0	1	0	0	0	0	0	1	0	0
שׁ	ʃ	0	0	0	0	0	0	0	1	0	0	0	1	0	0
שׂ	s	0	0	0	1	0	0	0	0	0	0	0	1	0	0
תּ	t	0	0	0	1	0	0	0	0	0	1	0	0	0	0
ת	t	0	0	0	1	0	0	0	0	0	1	0	0	0	0

F.   Korean

TABLE 13.   Visually systematic Korean consonants

Voiced		Voiceless		Tensed	
ㄱ	/ g /	ㅋ	/ k /	ㄲ	/ kₙ /
ㄷ	/ d /	ㅌ	/ t /	ㄸ	/ tₙ /
ㅂ	/ b /	ㅍ	/ p /	ㅃ	/ pₙ /
ㅅ	/ s /			ㅆ	/ sₙ /
ㅈ	/ dʒ /	ㅊ	/ tʃ /	ㅉ	/ tʃₙ /
ㅇ	/ ŋ /	ㅎ	/ h /		

TABLE 14.   Korean mono-thongs included in the study

Mono-thongs	
ㅏ	/ a /
ㅓ	/ ʌ /
ㅗ	/ o /
ㅜ	/ u /
ㅐ	/ e /
ㅔ	/ ɛ /
ㅚ	/ ø /
ㅟ	/ y /
ㅡ	/ ɰ /
ㅣ	/ i /

TABLE 15. 88 Korean fonts examined

굴림	꽃길	FB이철수80목판M
돋움	개미똥구멍	FB이철수90목판TM
바탕	신라장	FB이철수90목판M
궁서	제주한라산체	FB이철수2000목판TM
맑은고딕	제주고딕체	FB이철수2001목판M
나눔고딕	제주명조체	FB이철수2001목판TM
나눔명조	부산체	Yoon다정
나눔손글씨붓체	고양체	Yoon민홍
나눔손글씨 펜체	고양일산체	Yoon세희
나눔바른고딕	오성과한음체	Yoon아혜
나눔바른펜	약령리체	Yoon지영
나눔스퀘어	전라북도체	Yoon지희
나눔스퀘어라운드	푸른전남체	Yoon형오
Noto Sans CJK KR	KoPub돋움체	Yoon홍수
Noto Serif CJK KR	KoPub바탕체	감남윤
도현체	다온청년고딕	이현지
주아체	EBS주시경체	윤태민
한나는11살체	EBS훈민정음	한둥근체 돋움
간이벽온방	EBS훈민정음새론체	한둥근체 바탕
대한민국독립체	KBIZ한마음고딕	KCC은영
법정체	KBIZ한마음명조	KCC김훈
대한체	도서관체	한글누리
월인석보체	호국체	기령해랑체
고도체	간이벽온방	
아리따돋움	이롭게바탕체	
HS봄바람체2.0	tvN즐거운이야기체	
HS가을생각체	티몬몬소리체	
HS겨울눈꽃체	빙그레체	
HS두꺼비체	스웨거체	
가비아솔미체	한겨레결체	
가비아납작블럭체	조선일보명조체	
미생체	동그라미재단	
신비논일곱살	FB이철수80목판TM	

A New Approach to the Decipherment of Linear A, Stage 2

Cryptanalysis and Language Deciphering: A "Brute Force Attack" on an Undeciphered Writing System

Loh Jia Sheng Colin · Francesco Perono Cacciafoco

Abstract. This paper discusses the attempt of an algorithmic approach to contribute to the decipherment of Linear A. With the assistance of software developed in Python, Linear A clusters can be compared to various dictionaries of languages respecting a certain degree of chronological and geographical compatibility, as a "brute-force attack" for the reconstruction of new clusters of Linear A symbols.

1. Introduction

Linear A is a writing system of the Ancient Aegean Minoan Civilisation of Crete that was in use between approximately 1850 and 1450 BCE (Olivier, 1986). Linear B is a syllabic writing system partly derived from Linear A that was used to transcribe Mycenaean Greek. Both Linear A and Linear B have been discovered by British archaeologist Sir Arthur John Evans during excavations between 1886 and 1901 (Chadwick, 1967). The terms "Linear A" and "Linear B" were formulated by Evans based on the linear structure of inscriptions found on the tablets, in contrast to the pictographic writings that were used extensively during the same time period. Linear A samples were discovered to occur in

Loh Jia Sheng Colin   0000-0001-7009-5007
School of Physical and Mathematical Sciences,
Nanyang Technological University, Singapore
`jloh035@e.ntu.edu.sg`

Francesco Perono Cacciafoco   0000-0002-0977-063X
Linguistics and Multilingual Studies Programme,
School of Humanities,
Nanyang Technological University, Singapore
`fcacciafoco@ntu.edu.sg`

various locations such as Crete, Aegean islands of Kea, Kythera, Melos and Thera, and even mainland Greece (Olivier, 1986).

Currently, the Linear A corpus consists of approximately 1,400 artefacts with Linear A inscriptions, with signs appearing over 7,400 times. A large majority of these Linear A signs was used for administrative documentations found on tablets, roundels and seals (Schoep, 2002). Hypotheses regarding the origin of Linear A script and of the Minoan language include the Luwian Hypothesis, the Semitic Hypothesis and more. These will be discussed in later sections of this paper.

Despite numerous attempts by scholars and glyph-breakers, Linear A continues to remain undeciphered, as researchers' attempts to attribute a language family relation with Linear A have provided only a limited amount of meaningful results. This paper explains the use of a programme written in Python programming language (Eu, Xu, and Perono Cacciafoco, 2019) to isolate potential Linear A clusters for further analysis, with the intention of identifying potential family relations with Linear A and other dictionaries of languages.

2.   Selected Literature Review

Among the attempts made to decipher Linear A, many involve comparison with Linear B (Petrolito, Petrolito, Perono Cacciafoco, and Winterstein, 2015). Linear B, a syllabic writing system used in Crete, has been deciphered by the architect and philologist Michael Ventris in 1952, with the assistance of the linguist John Chadwick (Chadwick, 1967). They discovered that Linear B was used to transcribe Mycenaean Greek. Given that a large portion of Linear B was a derivation of Linear A, coupled with similarities between signs in both writing systems, there were reasonable justifications to support the conclusion of provisionally assigning Linear B phonetic values to Linear A signs that appear graphically similar. However, attempts to use Linear B directly for deciphering Linear A were unsuccessful. The results appear to be inconclusive as large amounts of 'meaningless' Linear A words were generated (Godart, 1984). As such, the direct use of Linear B alone might prove to be insufficient to decipher Linear A, thus motivating the studies of a variety of languages that are chronologically and geographically mutually compatible as possible relations to Linear A.

There have been studies carried out to decipher Linear A through potential association of Linear A with other language families. Vladimir Ivanov Georgiev believed that Linear A has connections with Greek. Based on his work published in 1952, Georgiev speculated that the Linear A inscriptions found on Haghia Triada tablets were transcribing Greek. Several studies have also highlighted that Linear A was associated with the Indo-European language family. Studies by Gareth A.

Owens (1999) have postulated the possibility that Linear A might belong to the Indo-European language family with relations to Greek, Sanskrit and Latin. Leonard R. Palmer (1961) proposed that Linear A could be an Anatolian language, possibly Luwian. Palmer posited his hypothesis due to a possible historical event which indicates the invasion of Crete and Greece by Indo-European peoples around the second millennium BC (ibid.), necessitating the migration over to Crete. Similarly, Gregory Nagy (1963) proposed a hypothesis that Linear A was largely similar to Luwian, a language that belongs to the Anatolian branch of the Indo-European language family.

Theories that propose a relationship between Linear A and Luwian have been constantly challenged and remain controversial in the academic community. Critics have indicated that Palmer's work is highly dependent on the interpretations of the Linear A tablets that can entail varied interpretations attributed to a limited understanding of Linear A orthography.

Linear A has also been associated with the Semitic language family. Semitic languages originate from the East of the Mediterranean and are speculated to be in use from around 3800 to 3500 BCE (Olivier, 1986). Based on chronological and geographical facts, Linear A is regarded to be compatible with the Semitic language family. Cyrus H. Gordon (1982), a scholar with an extensive knowledge of Semitic languages, was one of the first to attempt deciphering Linear A through comparison with the Semitic language family. By assigning Linear B phonetic values onto Linear A signs, Gordon was able to identify certain words present in Linear A that appeared to be largely similar to that of those belonging to the Semitic language family, such as Hebrew and Akkadian. Thus, he drew the conclusion that Linear A might be connected to Semitic languages, and West Semitic ones, in particular.

In support of the Semitic Hypothesis, archaeological evidence reveals trade practices by the Minoans over Semitic-speaking languages like Eastern Mediterranean and findings of Minoan-style artefacts discovered in areas that include Cyprus and Canaan (Bradley, 2014). This evidence supports the hypothesis that there was language contact between the language of Linear A and Semitic language family.

Nevertheless, many scholars continue to remain sceptical about Linear A having a language contact with the Semitic language family. Sceptics have argued that Gordon's work on the comparison of Linear A and the Semitic language family has major flaws. Firstly, the matches which Gordon has identified appear to be mainly vocabulary terms. Next, Gordon has adopted a methodology of associating elements from various Semitic languages, like Akkadian and Canaanite, for comparison with Linear A. However, the act of doing so has suggested that the Semitic hypothesis might not be conclusive due to the lack of linguistic evidence.

While some scholars continue to argue against Semitic being a possible family for the language represented by Linear A script, there are still recent studies on the decipherment of Linear A suggesting a possible Semitic connection. Eu, Perono Cacciafoco, and Cavallaro (2019) studied Linear A libation tables, in an effort to identify Semitic roots present in recurrences found in these tables. Eu Min et al. obtained very limited and sporadic matches that do present evidence for the relation of Linear A to the Semitic language family (Eu, Perono Cacciafoco, and Cavallaro, 2019).

The small number of such matches could be attributed to the small sample size of Linear A artefacts to study on, as well as the possibility that Linear A was used to encode a ritual language, involving a more complicated writing style.

American scholar John G. Younger set up a Web application[1] in 2000 dedicated to his work on Linear A, in an effort to provide better access to Linear A resources to other scholars. With the use of Linear B Syllabary, Younger has managed to transcribe most Linear A signs. Furthermore, Younger has attempted a reconstruction of Linear A based on the analytical interpretation of possible shared symbols between Linear A and Linear B. Through such forms of development, Younger was able to locate possible toponyms in Linear A that are pre-Greek but compatible with Linear B transcriptions.

Apart from the above-mentioned hypotheses, there are also other possible connections proposed by scholars. Margalit Finkelberg (2001) compared the phonological and morphological profiles of Minoan with those of other languages, in order to narrow down the range of possible languages associated with Linear A. The morphological profile of Minoan was compared to Greek, Lydian, Hittite, Luwian and Lycian. This comparison was based on the preliminary readings of certain Minoan texts. Consequently, Finkelberg proposed that Linear A might be an ancestor of Lycian or possibly a distant relative of it.

In recent years, studies on Linear A as well as the decipherment attempts made by scholars have included an algorithmic approach, with the introduction of high-computational power. Perono Cacciafoco (2017) has proposed the possibility of having the Linear A inscriptions analysed beyond the grammatical level, but through comparisons with other languages and language families according to the Linear A grammatological elements. This novel approach of deciphering Linear A could result in the reconstruction and recombination of new clusters of Linear A, based on the languages compared with the Minoan language. Peter Z. Revesez (2017) has proposed the language of Linear A to be connected with the Uralic language family. In his study, Revesez has introduced an algorithm to obtain the 'syllabic values of Linear A

---

1. http://people.ku.edu/~jyounger/LinearA/.

signs'. With these 'syllabic values', Revesez constructed a dictionary of Uralic-Minoan language that translates Linear A documents from the corpus: published by Louis Godart and Jean-Pierre Olivier in the 70s and 80s of the 20th century, GORILA is a Linear A corpus containing Linear A inscriptions. GORILA contains five volumes (Godart and Olivier, 1976a,b; 1979; 1982; 1985) and was made available digitally on the Web in the 21st century[2]. Like many previous attempts of deciphering Linear A, Revesez's work was heavily scrutinised as cross-family comparisons were not conducted, and the pertinent issue of varied interpretations of the 'syllabic values of Linear A' was not properly addressed in Revesez's work.

While it may appear that deciphering Linear A is an insurmountable challenge, it is often beneficial to consider other perspectives in approaching such problems (Tan, 2018), as an alternative means to attain a solution.

3. Methodology and Preparation of Materials

In contrast to past attempts to decipher Linear A, a comprehensive analysis and comparison of Linear A clusters should be conducted beyond the grammatological level, through methods originating in cryptanalysis. By analysing the combinatory data and comparison frequencies with language families of different natures, new possible clusters of Linear A words can be reconstructed and recombined. This project adopts and expands on the method of a "brute-force attack" to attempt deciphering Linear A, through the use of a programme developed in Python programming language (Eu, Xu, and Perono Cacciafoco, 2019).

For the comparison of Linear A with other languages, a set of documents containing dictionaries of various languages and language families has been generated digitally as input into the programme functions. These documents have data stored in spreadsheet files in order to be easily editable by a human. First, a compiled master list of transcribed Linear A words has been created out of the collections of text and Linear A samples contained in Olivier and Godart's Corpus of Inscriptions in Linear A, namely GORILA vols. 1 to 5 (Godart and Olivier, 1976a,b; 1979; 1982; 1985). The transcription of Linear A words has involved assigning phonetic values of Linear B symbols that appear compatible with the Linear A characters. Symbols in Linear A that do not resemble graphically to any of the symbols present in the Linear B syllabary have been replaced by three-digit numbers. For example, in Haghia Triada tablet HT1, there is a Linear A word transcribed by KU-[AB056]-NU, in

---

2. `http://mnamon.sns.it/index.php?page=Risorse&id=19.`

which the three-digit number '056' is used to denote that unique Linear A symbol (see Fig. 1, where KU, AB056 and NU correspond to ⸢, ⊟ and ⊦ resp.).

FIGURE 1. Pictorial representation for Linear A transcriptions of the HT1 tablet. Source: GORILA vol. 1 (Godart and Olivier, 1976a, p. 3)

A newly reconstructed digital corpus of Linear A signs is used to complement the database of Linear A signs and symbols present in GORILA vols. 1 to 5 (Godart and Olivier, 1976a,b; 1979; 1982; 1985). This allows more rigorous statistical analysis of Linear A signs and sign sequences, as comparative analysis will become much more efficient as compared to the printed version of Godart and Olivier's corpus.

In the future, dictionaries of languages from different language families will be required for comparison with the Linear A clusters. Examples of such dictionaries would include Luwian, Anatolian, Hamito-Semitic and Hittite.

4. Overview of the Programme Functions

Our programme aims at the implementation of a "brute-force attack" method as an attempt to decipher Linear A. In the context of cryptography and cryptanalysis, a "brute-force attack" is defined to be "an exhaustive search method in order to recover the secret key in a cryptosystem by testing all possible combinations" (Verdult, 2015). Through clever

optimisations and after a sufficient amount of computation, we aim to associate a possible language family to the Linear A writing system. The "brute-force attack" on the Linear A symbols can help in constructing phoneme *n*-gram databases with respect to the languages of comparison. Consequently, the Python programme aims to develop an automatic procedure for evaluating language sources that would be of "best fit" to Linear A. For the development of the programme, the 'pandas' and 'PyQt' Python modules have been used extensively for the purpose of data analytics and the creation of a graphical user interface (GUI) for the programme. In particular, the programme can be segmented into two portions—the specific decipherment approach (§ 4.3) and the general decipherment approach (§ 4.1). Aside from the aforementioned two portions, we have also incorporated the use of the Linear A fonts into the study of potential clusters of Linear A words found in various Linear A tablets.

4.1. General Decipherment (GD) Approach

In the GD function, users can input any spreadsheet file available on their computers, provided data are stored only in the first column of the spreadsheet file and the file is in CSV (comma-separated values) format. Using the pandas module, the programme carries out similarity comparisons between the words present in the uploaded file with those of the Linear A master list. Results generated through the GD function are displayed in a clear table format consisting of four columns, titled "Identical Matches," "Linear A word," "Original Word" and "Source". An example can be seen in Figure 2, in which there are 19 matches obtained from the programme's GD function. In particular, there are 6 matches for character "r," each of which originates from a different source, including ZA011b, HT27b and HT85b.

The results of word comparison can be downloaded locally as CSV files. Allowing an indiscriminate comparison between words from various language dictionaries and lexical lists and the Linear A master list, we can provide a large-scale brute-force attack on the current Linear A corpus while the decipherment of Linear A would involve a more statistical approach.

4.2. Modifications to GD

A modified version of GD, named "General Decipherment 2" (GD 2), will be implemented in the programme. By using the GD 2 function, users will be able to make dynamic changes in their comparisons and alter, not only the dictionary list, but also the Linear A master list—these
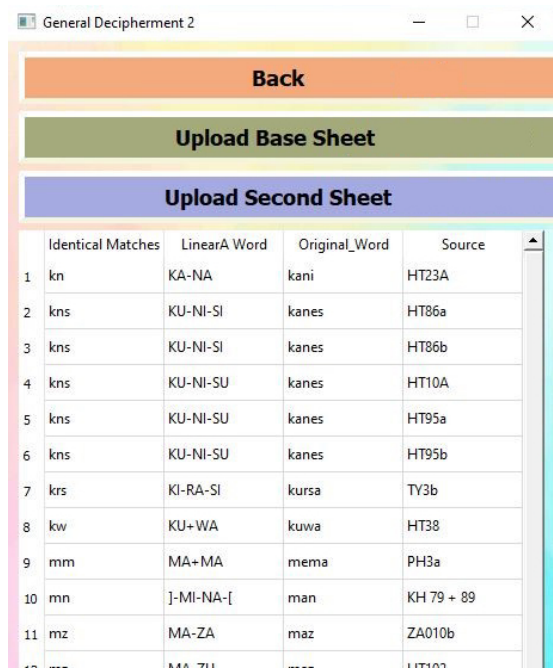
FIGURE 2. GD results of an unknown dictionary

two will be referenced to as "Comparison Sheet" and "Base Sheet" respectively. This form of comparison will be more beneficial in carrying out a more comprehensive analysis on the clusters of Linear A words, by incorporating the findings in the frequency analysis of the 3-digit numbers available in the SD portion of the programme. The frequency analysis of these numbers will be discussed in later sections of this paper. The only restrictions on the "Base Sheet" is that it is stored in CSV format. In the CSV file, the programme only uses the data stored in the first three columns: "Source," "New Format" and "Linear A word". Results obtained from this function will be available for download in CSV format. An example of the results obtained from GD 2 can be seen in Figure 3, where there are 5 matches for the string "kns," obtained from

the programme's GD 2 function. These matches originate from different sources, including HT86a, HT86b and HT10A.

FIGURE 3. GD 2 results of an unknown dictionary and modified Linear A master list

Allowing the users with dual choices of file input into the programme will provide a higher degree of freedom for more efficient research and analysis to be conducted on any possible cluster of Linear A words.

4.3. Specific Decipherment (SD) Approach

Under the SD function, numerous dictionaries from different language families are incorporated into the programme for comparison with the Linear A words present in the master list. Using a 'consonantal' approach for analysis, in which vowels occurring in words in the dictionaries will be provisionally removed, the programme will carry out a search in the master list to have a one-to-one character match between the Linear A clusters and the modified strings in the dictionaries.

This comparison adopts a consonantal approach that has been introduced for use with the Semitic family languages and Afro-Asiatic lan-

Figure 4. Digital recreation of the Linear A tablet HT 95b. The Linear A word "KU-NI-SU" is circled in red (Perono Cacciafoco and Cavallaro, 2020)

guages such as Ancient Egyptian. The approach has also been adopted for comparison for Indo-European languages. This is because consonantal clusters are more stable and consequently it is easier to highlight morphological parts of lexical items of a language and of their roots. In this aspect of the programme, identical character matches will be collated and displayed in a clear table format with the same four columns as the GD functions. An example of the result can be seen in Figure 5. Through such a comparison, dictionaries with high incidence of similarities can be quickly identified and isolated for further analysis.

As shown in Figure 5, the programme reads the results in four columns. Upon a one-to-one character match between the modified Luwian words as well as the Linear A clusters, the first column would output the results under 'Identical Matches'. According to the results, the SD function would select the original Luwian word, Linear A words as well as the source of the Linear A word. These would be available in the next three columns, respectively.

Aside from the direct comparison with words between Linear A and dictionaries, the programme function also conducts a series of frequency analyses. By allowing the 'triplets' present in Linear A clusters to use any possible character ranging from A to Z, the programme will analyse the frequency in which a character replacing the 3-digit number would achieve a one-to-one character match. An example of the results of the analysis of 3-digit numbers can be seen in Figure 6.

In Figure 7, we have the analysis of the triplets' one-to-one correspondence with the characters ranging from A to Z. It indicates that between the Thracian dictionary and Linear A master list, the triplet '029' has a one-to-one match with <a> one time, <b> three times, <c> five times, etc.

FIGURE 5. SD comparison results between Luwian dictionary and Linear A master list

FIGURE 6. Digital recreation of the Linear A tablet HT 27b. The Linear A cluster "RE" is circled in red (Perono Cacciafoco and Cavallaro, 2020)

The study of such frequency analyses can serve to justify the replacement of the 'triplet' with the character of the highest frequency of identical match. This process enables the reconstruction of possible new clusters of Linear A words, beyond the available clusters from inscriptions.

	1	2	3	4	5	6
1	Triplets	a	b	c	d	e	f
2	056	2	1	0	0	0	1
3	021	1	0	2	0	0	1
4	329	0	0	0	0	1	0
5	301	3	4	4	0	0	3
6	029	1	3	5	0	0	1
7	188	1	3	1	1	0	0
8	076	1	3	3	2	2	2
9	123	1	0	4	0	0	1
10	305	1	1	0	1	0	1
11	312	0	1	1	0	0	1
12	100	0	0	1	1	0	0

FIGURE 7. SD 3-digit number analysis between a Thracian dictionary and the Linear A master list

4.4.   Introduction of Linear A Fonts

The Linear A fonts introduced into the programme were created by Sabrina Soo Su-Ann, a fellow member in the Nanyang Technological University Linear A decipherment team. They expand the Unicode chart of Linear A, incorporating variants of Linear A signs.

The introduction of Linear A fonts will be represented by a function known as LinAfontConv in the programme. Through it, users can input their own files in Microsoft Excel Open XML Spreadsheet (XLSX) file format, provided they contain four columns labelled "Identical matches," "Linear A Word," "Original Word" and "Source" as shown in Figure 6. The function "LinAfontConv" will consider information under the "Linear A word" column, and display the output as the Linear A characters under another column titled "Linear A fonts". Similarly to the other functions in the programme, users can download the output file, which now contains the additional column "Linear A fonts," locally into their computers, in XLSX format.

Through such forms of linguistic analysis, we are able to identify and study potential clusters using the Linear A characters present in the tablet inscriptions. The incorporation of Linear A fonts will provide a better understanding of how Linear A was used by the Aegean Minoan people as a writing system when compared with various languages and language families.

1	Identical M	Linear A Word	Original Word	Source	Linear A fonts
2	pr	PARA	pari	HT128a	∓Ⲓͻ
3	pr	PARA	pari	PH3a	∓Ⲓͻ
4	pr	PARE	pari	HT 4	∓Ψ
5	pr	PURA	pari	HT116a	⅔Ⲓͻ
6	pr	PURE	pari	PK ZA 11	⅔Ψ
7	srr	SARARA	sarra	HT30	ΥⲒͻⲒͻ
8	srr	SIRERE	sarra	PH31a	ΨΨΨ

FIGURE 8. Results after input into "LinAfontConv" function

As seen in Figure 8, data from the input file should be ordered based on "Identical match," "Linear A word," "Original Word" and "Source". The programme will then output the data in the "Linear A word" column into Linear A fonts in the last column.

4.5. Preliminary Results and Discussion

With the assistance of the programme, we were able to attest identical matches between Linear A transcriptions and other dictionary lexical items. Figure 9 shows the examples of two consonantal clusters "PR" and "SR". The consonantal cluster of "PA" is derived from the Linear A cluster "PA-RE," located in HT4 of GORILA vol. 1 (Godart and Olivier, 1976a). Among the words from Hittite, Thracian, Anatolian, Luwian and Hamito-Semitic language families, with the same syllabic structure, a possible phonetic counterpart could be "PARI" from the Luwian dictionary (Melchert, 1993). The Luwian word "PARI" represents "forth or away" according to the dictionary. Similarly, "PA-RA" in Linear A has a possible counterpart in the Hamito-Semitic word "PARA" (Orel, 1994); "SI-RU" has a possible counterpart in the Thracian word "SIRIU" (Paliga, 2006); "SU-RE" with the Pre-Basque form "SUR" (Trask, 2008). Going through more examples of the consonantal clusters obtained by the programme, we hope to obtain a better understanding of what Linear A words can represent when compared with other language families.

5. Conclusion

Overall, to improve the efficacy of the Python programme, there will be only three functions available for users. These three functions are the aforementioned SD function, GD 2 function as well as the "LinAfontConv" function.

Consonantal Cluster		Linear A word		Dictionary Word
PR		PA-RE (𐙷𐘹)		Luwian: PARI
		Found in HT 4		(Defintion: forth, away)
PR		PA-RA (𐙷𐘷)		Hamito-Semitic: PARA
		Found in HT128a, PH3a		(Definition: Equid, Onager)
SR		SI-RU (𐘹𐘼)		Thracian: SIRIU
		Found in HT55a, TR ZA 1, IO ZA 15, PK ZA 10		(Definition: Vrancea region (Romania))
SR		SU-RE (𐘷𐘹)		Pre-Basque: SUR
		Found in HT 32		(Definition: Pour out)

FIGURE 9. Analysis of results obtained from the programme's SD function

FIGURE 10. Digital recreation of the Linear A tablet HT55a. The Linear A cluster "SI-RU" is circled in red (Perono Cacciafoco and Cavallaro, 2020)

The programme will be provided as an executable. Using the PyQt Python module, the programme will have a simple GUI to enable easy access to the functions. A shortcut will be created to enable quick distribution of the programme. This will also allow users to have direct access to the programme even without having the necessary dependencies and Python modules installed on their devices.

The Python programme has the potential to be further developed to enable possible image and text recognition as well as natural language processing through the use of the TensorFlow module. The incorporation of Linear A fonts in the programme can also allow the prediction of potential clusters of Linear A signs for further frequency analysis.

While the programme is able to conduct statistical and comparative analysis of dictionaries of various languages compared to Linear A, there are still certain areas of the programme functions that can be further improved.

Furthermore, with just the use of PyQt and pandas Python modules, the programme is unable to "read" and "render" the Linear A fonts. As a result, the programme will have to employ the use of an external software, such as Microsoft Excel or TEX, to ensure clear data rendering of the results using the Linear A fonts. As an improvement to this con-

straint, the Python module matplotlib will be introduced to render the Linear A fonts in the programme directly.

No doubt, the task of deciphering Linear A continues to be daunting to many researchers as the ancient Aegean Minoan writing system remains an unsolved enigma (Tan, 2018). The systematic and multidisciplinary approach consisting in using such a programme to decipher Linear A would be a shift away from the past philological attempts, as it involves methods from comparative linguistics and cryptolinguistics. This is a novel approach to tackle Linear A script and the elusive language behind it, and more research in this field will be able to be carried out.

A brute-force attack to the Linear A texts with our programme allows simultaneous comparison between Linear A and other languages from language families such as the Semitic and the Indo-European families, and provides us with a better understanding of the Minoan language. This could serve as a strong foundation for the crypto-linguistics approach of deciphering Linear A, where further work on Linear A can be carried out to facilitate the identification of a language family to which the Minoan language may belong.

6.   Acknowledgements

References

Bradley, P. (2014). *The Ancient World Transformed*. Cambridge: Cambridge University Press.

Chadwick, J. (1967). *The decipherment of linear B*. Cambridge: Cambridge University Press.

Eu, N., F. Perono Cacciafoco, and F. Cavallaro (2019). "Linear A Libation Tables: A Semitic Connection Explored." In: *Annals of the University of Craiova: Series Philology, Linguistics / Analele Universității Din Craiova: Seria Științe Filologice, Linguistică*. Vol. 41. 1/2, pp. 51–63.

Eu, N., D. Xu, and F. Perono Cacciafoco (2019). "Coding to Decipher Linear A." In: *Proceedings of the 2019 Pacific Neighbourhood Consortium Annual Conference and Joint Meetings (PNC)*. Singapore: Nanyang Technological University, pp. 44–48.

Finkelberg, M. (2001). "The Language of Linear A: Greek, Semitic, or Anatolian?" In: *Greater Anatolia and the Indo-Hittite Language Family*. Ed. by R. Drews. Vol. 38. Journal of Indo-European Monograph: Series, pp. 81–105.

Godart, L. (1984). "Du linéaire A au linéaire B." In: *Aux origines de l'hellénisme: La Crète et la Grèce*. Paris: Publications de la Sorbonne, pp. 121–128.

Godart, L. and J.-P. Olivier (1976a). *Recueil des inscriptions en linéaire A. 1. Tablettes éditées avant 1970*. Vol. 21. Études crétoises. Paris: Librairie orientaliste Paul Geuthner.

——— (1976b). *Recueil des inscriptions en linéaire A. 3. Tablettes, nodules et rondelles édités en 1975 et 1976*. Vol. 21.3. Études crétoises. Paris: Librairie orientaliste Paul Geuthner.

——— (1979). *Recueil des inscriptions en linéaire A. 2. Nodules, scellés et rondelles édités avant 1970*. Vol. 21.2. Études crétoises. Paris: Librairie orientaliste Paul Geuthner.

——— (1982). *Recueil des inscriptions en linéaire A. 4. Autres documents*. Vol. 21.4. Études crétoises. Paris: Librairie orientaliste Paul Geuthner.

——— (1985). *Recueil des inscriptions en linéaire A. 5. Addenda, corrigenda, concordances, index et planches des signes*. Vol. 21.5. Études crétoises. Paris: Librairie orientaliste Paul Geuthner.

Gordon, C. H. (1982). *Forgotten scripts: Their ongoing discovery and decipherment*. New York: Basic Books.

Melchert, Craig H. (1993). "Cuneiform Luvian Lexicon." In: vol. 2. Lexica Anatolica. Chapell Hill, NC: self-published.

Nagy, G. (1963). *Greek-like Elements in Linear A*. Durham: William H. Willis.

Olivier, J.-P. (1986). "Cretan writing in the second millennium BC." In: *World Archaeology* 17.3, pp. 377–389.

Orel, V. (1994). "On the ancient contacts between Hamito-Semitic and north Caucasian." In: *Folia Linguistica Historica* 15.1–2.

Owens, G. A. (1999). "The Structure of the Minoan Language." In: *J. Indo-Eur. Stud.* 27.1/2, pp. 15–56.

Paliga, Sorin (2006). *Etymological Lexicon of the Indigenous (Thracian) Elements in Romanian*. București: Editura Evenimentul.

Palmer, L. R. (1961). *Myceneans and Minoans: Aegean pre-history in the Light of the Linear B tablets*. Faber & Faber.

Perono Cacciafoco, F. (2017). "Linear A and Minoan: Some New Old Questions." In: *Annals of the University of Craiova: Series Philology, Linguistics / Analele Universității Din Craiova: Seria Științe Filologice, Linguistică*. Vol. 39. 1–2, pp. 154–170.

Perono Cacciafoco, F. and F. Cavallaro (2020). "Linear A Corpus." `https://blogs.ntu.edu.sg/linear-a/`.

Petrolito, T. et al. (2015). "Minoan linguistic resources: The Linear A Digital Corpus." In: *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities (LaTeCH), Beijing, China*, pp. 95–104.

Revesez, P. (2017). "Establishing the West-Ugric Language Family with Minoan, Hattic and Hungarian by a Decipherment of Linear A." In: *WSEAS Trans. Inf. Sci. Appl.* 14, pp. 306–355.

Schoep, I. (2002). "Social and Political Organisation on Crete in the Proto-palatial Period: The case of Middle Minoan II Malia." In: *J. Mediterr. Archaeol.* 15.1, pp. 101–132.

Tan, K. M. W. Y. (2018). *The Minoan Engima: Deciphering Linear A*. Tech. rep. `http://hdl.handle.net/10356/76547`. Singapore.

Trask, R.L. (2008). *Etymological Dictionary of Basque*. Sussex: University of Sussex.

Verdult, R. (2015). "The (in)security of proprietary cryptography." PhD thesis. Radboud University, The Netherlands and KU Leuven, Belgium.

SigLA: The Signs of Linear A

A Palæographical Database

Ester Salgarella · Simon Castellan

Abstract. We present a database of inscriptions written in the (still undeciphered) Linear A script of Bronze Age Greece. We aim at developing a systematic, exhaustive and user-friendly open access database of all Linear A inscriptions. Such a research tool is currently missing, and is essential in order to carry out statistical and palæographical analyses within the epigraphic corpus, only available in print form at the moment.

1. Introduction

This paper presents an interdisciplinary project blending linguistics and computer science and aiming at developing a systematic, exhaustive and user friendly open access database of all inscriptions known to date written in the Linear A script of Bronze Age Greece (ca. 1800–1450 BCE), to date still undeciphered (see § 2). Such a research tool is currently missing, and is highly desirable inasmuch as essential in order to carry out statistical and palæographic analyses within the epigraphic corpus, currently available in print form only. In fact, one of the hindrances to decipherment prospects is the current impossibility to carry out any meaningful linguistic statistical analysis and palæographic sign comparison covering the whole corpus of Linear A inscriptions due to the limited resources available. This is especially true with respect to research tools, as all material is only available in (cumbersome) print form. Collecting the Linear A inscriptions in a unified database is of paramount importance to be able to answer sophisticated palæographical and linguistic questions about the Linear A script as well as the language (Minoan) it

Ester Salgarella ⓘD 0000-0001-5091-5311
St John's College, Cambridge, UK
E-mail: es636@cam.ac.uk

Simon Castellan ⓘD 0000-0001-5886-5793
Inria, Univ. de Rennes, CNRS, IRISA, Rennes, France
E-mail: simon.castellan@inria.fr

encodes, which will help us reconstruct the socio-historical context of the Minoan civilisation.

The database will record and display for cross-search comparison: (i) *linguistic information*: contextual occurrences of signs, their frequency and position within a tablet, as well as individual sign-sequences (i.e., words) and their relative position and frequency within the whole corpus; (ii) *palæographic variation*: the way in which particular occurrences of signs are drawn on a contextual basis, and how signs vary from inscription to inscription (intra-site analysis), and from location to location (inter-site analysis).

Emphasis will be put on allowing users to see the material evidence (e.g., quickly see all occurrences of a sign or a word in a particular location), in order to ease palæographic analyses that have so far been done tediously by hand by perusing the print corpus of Linear A inscriptions (known as *GORILA* (Godart and Olivier, 1976−1985), see § 2). Having a digital approach here, where occurrences of signs can be easily compared is key for carrying out comparative analysis. This is greatly simplified by the use of a database, given the very little information we can retrieve solely from the laconic textual structure of the inscriptions as they are (characterised by a great many abbreviations which require a context-driven interpretation of the same signs and/or sign-sequences), as well as the overall poor evidence in terms of quantity and preservation.

In what follows, we will describe the current situation of the Linear A evidence, the state of art in the scholarship and, most crucially, the problems we faced when trying to combine linguistic and palæographic evidence together, as well as the solutions we came up with to develop the features of the database. A first version of the database is available at the address `https://sigla.phis.me`.

2. *Ab Antiquo*: The Linear A Script of Bronze Age Greece

Linear A is a logo-syllabic writing system used in the Bronze Age (ca. 1800−1450 BCE) primarily on Crete, but also sporadically in Mainland Greece and the Aegean islands (for a concise overview of Linear A in context see esp. Decorte, 2018; Tomas, 2010a, pp. 18−25; more comprehensive studies are Davis, 2014; Salgarella, 2020; Schoep, 2002). Linear A was used by the so-called 'Minoans' to write down their language, the 'Minoan' language indigenous to Crete. Despite this broader geographical area having been Greek-speaking from around the end of the Bronze Age until today, Minoan still resists decipherment as it does not seem to be related to any of the Indo-European languages so far known (most notably Greek), nor does it to Semitic ones (spoken in the neighbouring areas, esp. Egypt and the Levant) (for a recent and thorough lin-

guistic analysis of Linear A see Davis, 2013; 2014, pp. 156–278). Hence, Linear A remains to date one of the world's still undeciphered writing systems.

Notwithstanding, we are in a position to be able to at least 'read', although with an approximation, and to an extend to interpret inscriptions written in Linear A. This is because Linear A functioned as a template for the creation of Linear B, a writing system used on Crete and in Mainland Greece in the time-span ca. 1400–1190 BCE by the Greek-speaking 'Mycenaeans'. Linear B was successfully deciphered as an early form of Greek in 1952 (for a summary of the decipherment process see esp. Chadwick, 1967; Judson, 2017; Pope, 2008). A good number of signs of this 'Linear Script' (on this terminology see esp. Salgarella, 2019; 2020) show the same, or a highly comparable, graphic shape and are therefore called 'homomorphic signs'. It is argued (lastly Steele and Meissner, 2017) that some of these signs are also to be taken as 'homophones', i.e., having a similar phonetic value. Hence, by applying the homomorphy-homophony principle, the phonetic values we know for Linear B signs are retrospectively applied to Linear A homomorphic signs, allowing for an approximate reading of Linear A sign-sequences.

From a typological as well as functional standpoint, both Linear A and Linear B are logo-syllabic writing systems, meaning that they consist of two functional categories of signs: (i) *syllabograms*, i.e., phonetic signs representing syllables (only open syllables of the type: single Vowel, Consonant-Vowel or Consonant-Consonant-Vowel: e.g., *a*, *pa*, *nwa*); and (ii) *logograms* (or "ideograms," on terminology see esp. Thompson, 2010), i.e., signs standing for entire words or concepts. This sub-division, however, is more marked in Linear B than it is in Linear A, where a sign can behave either way based on context. The function performed by a sign is often inferable from its position in the inscription: logograms are placed at the end of an entry (after sign-sequences interpretable as words and before numerals). Context is here of considerable help, since most Linear A inscriptions, and almost the entire corpus of Linear B texts, consist of clay documents functioning as records of economic transactions used for the bookkeeping of the Palatial administrations of Late Bronze Age Crete (and Mainland Greece for Linear B). As such, these fall into the broader category of 'administrative documents'.

The most common type is the clay tablet, recording the flow of incoming and outgoing goods, which was used in both Linear A and Linear B administrative practice (esp. Tomas, 2006; 2010b; 2011b; 2017a,b). Moreover, each administrative system had a number of system-specific documents. These are: for Linear A, roundels (understood to have functioned as some sort of receipts), and sealings of different types (see esp. Bennet, 2008; Hallager, 1996, p. 10); for Linear B, labels, nodules and noduli (esp. Bennet, 2008, p. 17, Hallager, 2011, pp. 65–68; Tomas, 2017b). Unlike Linear B, whose use was restricted to administrative pur-

poses only, Linear A is also attested on a variety of other supports used
in different contexts, falling into the general label of 'non-administrative
documents'. These inscriptions are understood to be mostly religious in
nature (e.g., the 'libation formula', see esp. Karnava, 2016). At present
the database only contains administrative documents, more precisely
the Linear A tablets found at the most prominent sites on Crete. How-
ever, the long-term plan is to implement the database by adding all in-
scriptions recovered so far in order to make it as comprehensive and
exhaustive as possible.

3.   *Ab Initio*: Developing a New Tool

3.1.   The Standard Corpus of Linear A Inscriptions

At this point, one may wonder, where and how are the Linear A in-
scriptions available to examine? The extant evidence (both administra-
tive and non-administrative documents, on any supports) is presented
in the five volumes of GORILA (Godart and Olivier, 1976–1985), pub-
lished by Louis Godart and Jean-Pierre Olivier some 40 years ago. This
still remains the only corpus of Linear A inscriptions, solely available
in print form (although scans have recently been put online by publish-
ers). However, more evidence has been coming up since the publication
of the corpus, and has been published in individual articles (an addition
to GORILA is in preparation by Del Freo and Zurbrach, 2011). As it
stands, the corpus shows a black-and-white photograph of each docu-
ment, followed by a drawing and two transcriptions: the first transcrip-
tion is faithful to the original layout of the text (to ease sign identifica-
tion in their original position on the actual document), while the second
transcription shows standardised sign shapes along with a functional
arrangement of the text (for easier interpretation of the record).
     The corpus was a considerable achievement for the time, since it
made the evidence accessible to the academic community for the first
time, allowing scholars to reach an accurate and detailed interpretation
of all the the material since then unearthed. However, as it is, the cor-
pus does have limitations: first, it is not intended for a wide readership,
and is only accessible to those who already have a basic knowledge of
Linear A given that neither a transliteration nor a transnumeration of
the inscription is given, but only a transcription (Linear A signs are usu-
ally best known by their classification number, e.g., AB 60, allowing for
quicker retrieval in the standardised sign list). Therefore, in order to
read a text the sign shapes shown in the transcription have to be checked
against the standardised list of Linear A signs (available at the begin-
ning of Volume 5, pp. xxii-xxvii): Linear A is composed of some 180
simple signs (representing a graphic and phonologic unit); and some
164 complex (or composite) signs (which are the combination of two or

more simple signs), on top of these there are some 30 fractional signs. A quite reliable transliteration of the texts, although subjective in places, is given by Younger (2000) based on Godart and Olivier (1976–1985) transcriptions. This contribution has so far proved to be useful, especially for linguistic analyses; however, a mere transliteration leaves out palæographical information. Second, another limitation of the corpus is its very format: a printed edition does not allow to carry out any statistical and comparative analysis of signs and sign-sequences. This resulted in slowing down comparative linguistic and palæographical research (unless one painstakingly collects their own dataset). A digital approach, therefore, is clearly needed to make the most of the evidence and promote further linguistic and palæographical research, allowing for complex searches. In fact, we may want to see which variant of a sign is used on a given document, how frequent such a variant is within the whole corpus of inscriptions or within a selected set of documents (e.g., site-specific or document-type-specific analyses), which variant distribution patterns can be observed, or to simply have an overall appreciation of the palæographical features characterising the Linear A evidence coming from a given find-place.

3.2. A Digital Approach: Challenges and Solutions

Turning the printed corpus into a digital database raised a few challenges. The first and more important challenge is that of copyright, as the images included in the corpus are not free to use. To circumvent the copyright issues surrounding the original drawings we decided to make our own drawings of each document, based on the standard corpus of inscriptions and as faithful to the originals as possible. This long process turned out to be fruitful, as it allowed us to separate distinctly each and every sign drawn on the tablet surface, to classify each sign individually and to mark its position within the inscription. As a result, drawings can be annotated with information that would not have been possible to extract automatically for comparative purposes. As an example, our database includes for display and analysis epigraphical features such as erasures (see § 4). To make the drawing process as smooth as possible, we opted to use Krita, a graphics editor, and to turn the corpus images into multi-layered images where each sign belongs in a different layer. Basic metadata can be encoded in these files, therefore our digital corpus becomes a set of Krita files, one per document of the corpus.

Another problem we had to face during the design of the database was classifying the data. Because of the nature of evidence, there is a lot of uncertainty: uncertain readings, unknown word boundaries, uncertain function performed by signs in isolation (e.g., logograms or transaction-signs?). Moreover, the standard terminology used in the scholarship is

itself ambiguous to some extent, and part of the work of designing the database was to resolve such ambiguity. For instance, the word *sign* can refer to the standardised shape of a sign (e.g., AB 01) or to a particular occurrence of that standardised sign on a particular document, or to a graphic variant of the sign (at times difficult to recognise as such). Digitalising the corpus forced us to impose the strict inflexibility of formal languages onto the flexibility of natural languages, and forced us to make some choices in cases where the evidence is not clear.

For all these reasons, we decided to develop our own software to deal with the database. The software has two main components:

– *Import.* This component turns the corpus of Krita files into a JSON database, and produces image files for each document, and for each sign attestation.
– *Interface.* We then developed a web interface, written in OCaml and compiled in JavaScript that entirely runs in the browser. The interface allows to visualise and search the data, and is presented in § 4.

We choose this architecture to favour simplicity and openness. The database is easily accessible and usable by other people outside the interface if so they wished, and JSON is one of the best supported data description languages. Moreover, the website can be downloaded and run locally. This also ensures a very small load on the server that does not run any computation and ensures that SigLA (or copies thereof) can be easily hosted. The main trade-off is performance as this is much slower than a relational database would be. We believe this is not a problem as the evidence for Linear A is relatively small (less than two thousands documents).

One other challenge in developing the interface was to allow flexibility in the queries that can be expressed, while still remaining user-friendly and simple for the most frequent queries. For the expressiveness part, we decided to represent queries as typed λ-terms, which are functional programs, built on primitive terms representing properties of the objects manipulated by the database. For instance, types include `Word`, `Sign`, `Document`, and properties include `words`, of type `Document →` `Word List`. Using types, we built an interface allowing the user to build the λ-term incrementally by showing them the possible properties that are available at any point in the query (see § 4.2). This proved to be expressive and easy to extend by adding more primitives.

4.   *Ad Hoc*: Main Features

4.1.   Visualising the Data

The first feature that SigLA offers is the possibility to inspect individual documents of the corpus along with their metadata. Metadata include: find-place, document typology (clay tablet, roundel, …) and dimensions,

density of information on the writing surface (total number of signs, total number of words). On the document displayed, individual sign occurrences are highlighted in different colours for ease of reference: when hovering with the mouse on a particular sign, information about it is displayed, such as its transnumeration (i.e., its classification number as standardly set out in Godart and Olivier, 1976–1985, Vol. V, pp. xxii-xxvii), its possible transliteration (i.e., approximate phonetic value), its function on the tablet (syllabogram, logogram, transaction-sign, fraction). Each sign is coloured according to its function: shades of blue for syllabograms (phonetic signs which are part of a word), green for logograms (more or less pictographic signs which stand for entire words or concepts), orange for fractions (fractional signs accompanying numbers), yellow for transaction signs (individual signs occurring in isolation usually on top of a tablet with a word divider on either side; see Schoep, 2002, pp. 39, 135, 140; Salgarella, 2020, pp. 50–54); and red for erasures (instances where the traces of a previously cancelled signs are still visible on the writing surface). However, at times the precise function a sign performs in a given context is unclear. In such cases, we decided to allocate the sign the function that it is most likely to perform based on context, but this choice may well be subject to revision. Moreover, we have come across a number of unclassified signs (in GORILA these are referred to with a question mark in the context where they appear and are not included in the standardised sign list), which we have labelled as such and are searchable in the database for further contextual analysis. In SigLA we use the question-mark for signs of doubtful reading or unreadable (instances where traces of a sign are visible, but the sign can not be recognised). Also in this case, all instances of unreadable signs across the whole corpus can be viewed.

The visualisation of the data as described above is displayed when viewing the document in *sign view,* illustrated in fig. 1. In addition to this setting, the document is also available in *word view,* showing coherent sign-sequences (words). Here only the syllabograms forming a word are highlighted (leaving aside logograms, transaction-signs and numerals): hovering on a sign selects the word it belongs to and clicking on such word allows the user to see other occurrences of the selected word across the corpus (and its relative position on each document for comparative purposes).

In *sign view* when clicking on the sign number, the user is redirected to the palæographical chart of the sign, which displays all occurrences of that sign across the corpus. This is one of the main goals of SigLA: to be able to compute automatically such charts, which are key to palæographical analysis and before had to be produced by hand by researchers. An example of such chart for sign AB 08 (phonetic value /a/) is displayed in fig. 2 (the figure only presents the chart relative to the site of Haghia Triada, but in SigLA all sites are available).

Document HT 23a

Switch to word view

- Site found: Haghia Triada
- Type of inscription: Tablet
- Number of signs: 30
- Number of words: 5
- Dimensions: 5.6 cm x 6.9 cm x 0.8 cm

Sign #22: qa

Function: syllabogram
Simple sign
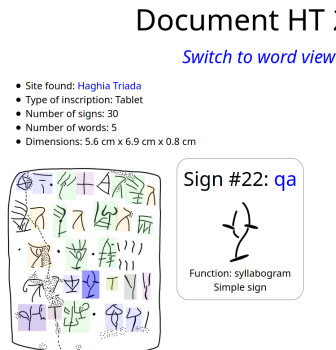
FIGURE 1. A view of a document in SigLA (*sign view*)

SigLA also comes with a sign list that displays all signs occurring in the corpus, again following the accepted sign classification set out in Godart and Olivier, 1976–1985, Vol. V, pp. xxii-xxvii. However, unlike in Godart and Olivier (ibid.), the sign list of SigLA does not use a standardised (hence, somewhat abstract) shape for each sign, but rather a particular occurrence of such sign that has been considered as representative by the authors. In the case of composite signs, their decomposition into simple signs (individual constitutive components) is also displayed, following the interpretation proposed in Godart and Olivier (ibid.) and with a refined notation introduced by the first author (Salgarella, 2020, pp. 54–59).

As last remark, we also decided to add a *Map of sites* in the Homepage of SigLA, showing all sites that have yielded Linear A evidence (both administrative and non-administrative). This is a reference tool that shall help users to locate sites on Crete, as well as to evaluate at first sight the distribution of find-places on the island.

4.2.   Searching the Database

SigLA allows users to search the corpus by providing three types of searches: (i) *sign search*: search for sign occurrences, (ii) *word search*: search for sign-sequences, (iii) *document search*: search for specific documents. SigLA supports searches of signs (simple or in composition) or words. Some examples are given below.

– *Sign search*: The sign search function allows users to look for a sign of their choosing either across the whole Linear A corpus or within a customisable subset of evidence. The end result of such search could be either palæographical charts showing all the occurrences of the
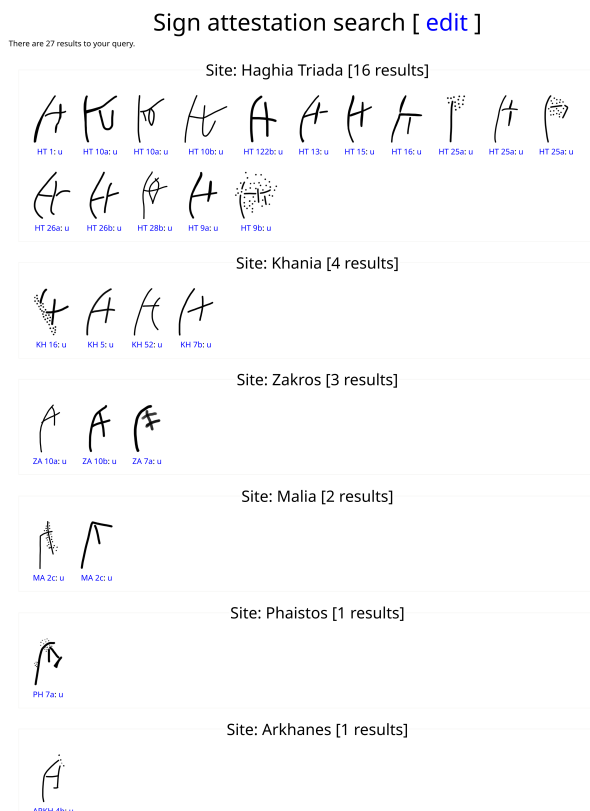
FIGURE 2. Palæographical chart for sign AB 10 (/u/) in SigLA

sign sought for in isolation for comparative purposes (as in fig. 2), or its contextual occurrence and position on complete documents with the occurrence of the sign highlighted. This search is of particular importance for the evaluation of sign frequency and use across sites and for assessing palæographical variation.

A similar search can be carried out with respect to erasures, which are here treated as signs. It is possible to run a search for erasures in order to assess their frequency and contextual occurrences (at times it is possible to understand the reason that led to the cancellation of a previously written sign). The result of such search is illustrated in (fig. 5), showing the erasures attested on a set of documents from Haghia Triada.

Finally, in *sign search* it is also possible to look for simple signs in combination: more precisely, for all attestations of a simple sign when in combination with other ones to produce composite signs (in Lin-

ear A a simple sign can be combined with multiple others). By way of example, let us take simple sign A 302 (𐘿): by using the nomenclature `A 302+` in the sign search box and running a search, we can view all attestations of composite signs having A 302 as one of their constitutive components (fig. 4).

– *Word search*: Let us assume we wanted to view all attestations of the word *ku-ro* (understood to mean 'total') within the Linear A corpus. By running a word search we end up seeing all contextual occurrences of *ku-ro*, as illustrated in fig. 3, showing all attestations on the documents from Haghia Triada (but in the database all sites are displayed). This viewing setting is also useful to carry out comparative analysis of the position of the same sign-sequence over different documents so as to get insights into the meaning of Linear A words (our knowing the meaning of *ku-ro* is more of an exception than the rule).

– *Document search*: The document search option allows to look for a particular document or a set of documents (customisable by the user) within the whole corpus or within a given site (or a selected combination of multiple sites). It is also possible to narrow down the search to a specific document type (e.g., tablet, roundel, label, etc.) so as to evaluate its frequency and distribution across sites. The end result of this search is the viewing of entire documents, which are displayed without additional highlight on specific features. Given that the metadata included in SigLA also contain information about document dimensions, this search allows to see and evaluate at first sight the relative proportions and sizes of all the documents within the corpus, allowing for comparative analysis of their sizes (as well as some pinacological features).

– *Quick search*: Finally, for easier searches, a quick search option is also available (displayed on the top bar), which can be used to quickly jump to a particular document (e.g., HT 12), sign (e.g., AB 60), or location (e.g., Knossos), without engaging with any of the aforementioned search interfaces.

These simple searches offer already a lot of improvement on the print corpus of Linear A inscriptions. However, as explained in § 3.2, SigLA also offers a number of more complex and advanced searches. Such searches are done by supplying a list of criteria that must all be met by the objects sought after. Such criteria are expressed using properties of the objects and can be quite sophisticated. In fig. 8, we show how one would enter the query "Search for documents in Haghia Triada that have a word of length greater than five". This search has two criteria (location and existence of such a word). The search query is composed interactively, and the user is guided at each step by viewing what are the possible properties they can use in the query: fig. 6 and fig. 7 depict the steps in this interactive process, where at each step we are offered the list of possible properties to use. The first step (fig. 6) is to select the

Site: Haghia Triada [12 results]

HT 119: ku-ro

HT 11a: ku-ro

HT 11b: ku-ro

HT 127b: ku-ro

HT 127b: ku-ro

HT 13: ku-ro

HT 25b: ku-ro

HT 25b: ku-ro

HT 27a: ku-ro

HT 89: ku-ro

HT 94a: ku-ro

HT 9b: ku-ro

FIGURE 3. Sign search: Attestations of *ku-ro* at Haghia Triada
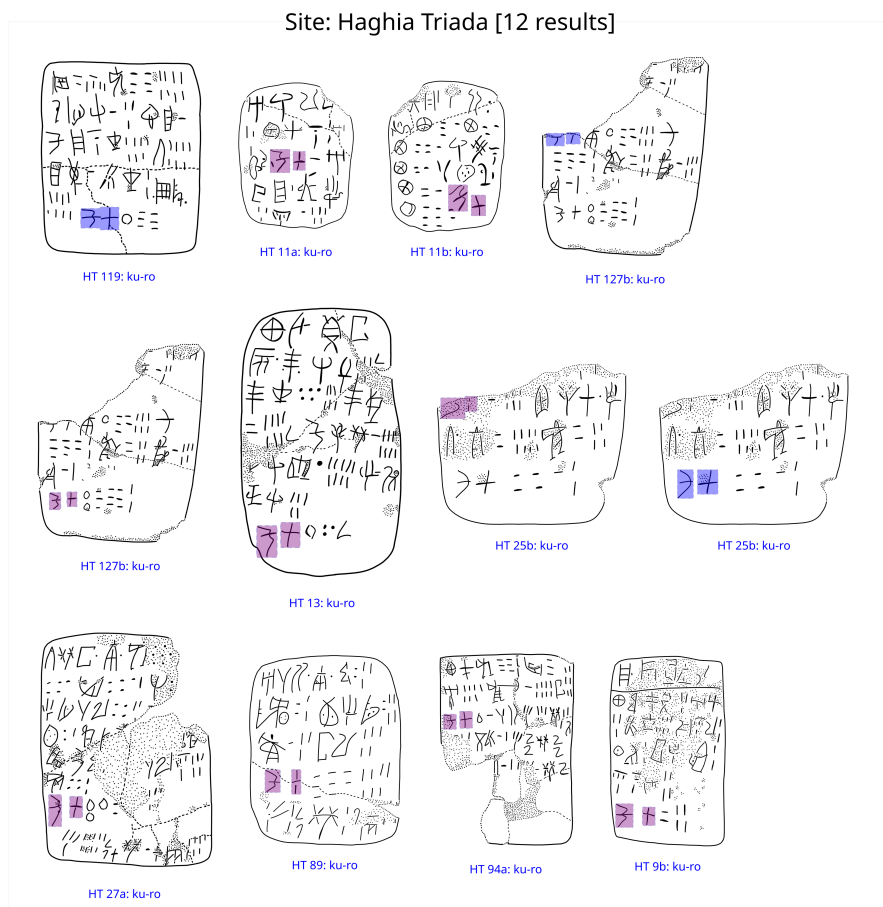
condition on documents *Contains one word satisfying*. The second step is to specify which words we are interested in, i.e., specify a predicate on words. Hence we have access to a different set of blocks, including *Word length* which is what we want (fig. 7). The third step is to specify that we want the length to be *greater than* a specific number, here five (fig. 8).

As also shown in fig. 8, results can be grouped and sorted in arbitrary ways, the default option being to sort them by site. Groups can also be nested (for instance to group results by site, and to group the result within each site by document).
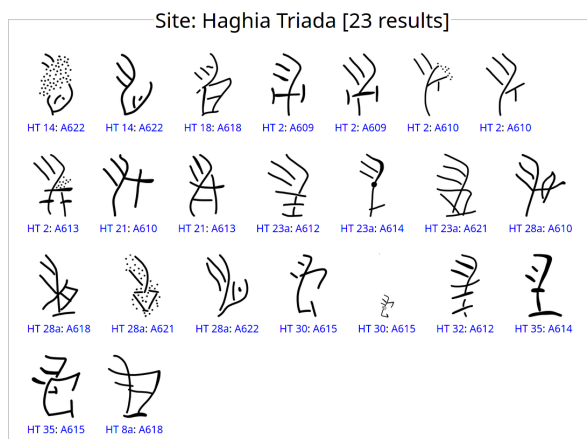
FIGURE 4. Attestations of complex signs containing A 302

5. *Ad Maiora*: Future Improvements

As it stands, the database is still under construction, although all the main features have already been developed in this first version. We are currently working towards implementing the database with a number of additional features, described in what follows.

In order to ease reading and interpretation of the Linear A documents displayed in the database, we are planning on adding a full transliteration of each inscription by using (approximate) phonetic values (based on comparison with Linear B, see discussion in § 2). At present such transliterations have only been made available by John Younger in his website 'Linear A Texts and Inscriptions in Phonetic Transcription and Commentary' (Younger, 2000). Younger's transliterations are based on Godart and Olivier (1976–1985), and are often improved with his own readings. However, Younger's website is mostly concerned with Linear A texts, much less so with the physical appearance of documents. Hence, palæographical features are not displayed. We hope that in this respect SigLA will complement, as well as integrate, Younger's work. In addition to the phonetic transliteration of Linear A inscriptions, we would also like to show a further transcription in Unicode characters. This feature may be of particular use to those who are less familiar with Linear A signs and their palæographical variation. In fact, Unicode characters reflect the standardised shapes of signs (and combinations thereof) as listed in ibid., Vol. V, pp. xxii-xxvii), and will ease legibility of the original inscription as preserved on the document surface. Adding this information will enhance clarity of reading and interpretation, as users will be able to recognise at first sight which signs are shown on a given document and to ap-

Figure 5. Erasures at Khania

preciate their palæographical variation as contextual occurrences. In this respect, SigLA presents itself as a didactic tool.

Another useful implementation will be the addition of a section for notes or comments after each document. This section shall accommodate information that will ease interpretation of inscriptions and texts, given that in most cases this process is problematic to say the least (due to the fragmentary state of preservation of a good many documents, our imprecise knowledge of the meaning of sign-sequences as well as the low frequency of cross-site sign-sequences, the multifunctionality of signs based on context, etc.). In this way, users will be guided to make sense of the texts and transcriptions as they appear by using their own judgment. Moreover, whenever possible, references will be made to individual studies dedicated to each document displayed or of interest for its interpretation. In fact, at present there is no 'Handbook' of Linear A, where to learn all the basics to interpret an inscription and situate it in its broader archaeological and historical context.

The addition of photographs of original documents remains a most cherished *desideratum* for the time being, as contingent upon copyright

FIGURE 6. Properties for document

FIGURE 7. Properties for words

permissions. We wish SigLA (or any alternative similar database) will one day host high-resolution (ideally 3D) images of Linear A documents for more thorough and accurate first-hand inspection of the documents' palæographical and pinacological features.

FIGURE 8. Example of a complex document search

6. *Ad Aeterna*: New Reseach Pathways

To conclude, we would like to pinpoint some of the potential applications and future pathways of research SigLA will allow.

First, it shall be possible to refine the Linear A sign repertory, by being able to clearly differentiate between signs and their variants and at the same time reaching a better global understanding of the structural characteristics of the writing system. This will ultimately lead to the appreciation of how many signs do represent the core of the writing system (in this respect see also Salgarella, in preparation).

Second, SigLA shall allow to carry out systematic research into scribal activity, resulting in a more thorough and reliable identification of scribal hands, at present still not clearly identified nor identifiable (in Godart and Olivier, 1976–1985, Vol. V, pp. 83–113, Godart put forward possible scribal hand attributions, but without explaining the reasons of his choices; on scribal hand identification see esp. Militello, 1989; Raison and Pope, 1971; Tomas, 2011a), as well as getting an idea of the overall number of scribes at work at a given site. By consequence, we should also be able to come to identify possible scribal/writing 'traditions' and their spread across time and space. This, in turn, will allow to throw light on matters pertaining to the acquisition and transmission of the writing system as well as writing practice, enabling us to assess the extent of literacy and the level of specialisation in writing and administrative practices in Minoan Crete. These are highly debated questions that scholars have already been addressing, exploring and trying to answer

with respect to Linear B. However, we are severely lagging behind with respect to Linear A.

Third, in the long term the database will also display information about the physical appearance and manufacture of the Linear A documents included, such as erasures (already searchable on SigLA), presence or absence of ruling lines to guide writing, presence of word-dividers, density of writing on the writing surface as well as textual arrangement, size comparison, presence of cuttings (if a tablet was cut after being inscribed). All this data will give us key information about the tablet manufacture process as well as the editing of the text itself.

Last but not the least, all these features, taken all together, shall enable us to carry out a more sophisticated and thorough palæographical comparison for the evaluation of (the degree of) palæographical similarities and differences both across sites and within a site, with the possibility for each and every user to narrow down their analysis to specific features by customising their own dataset based on their very own research interests.

This unique research tool, and the ensuing pathways it shall enable, will help shed light on all above areas of investigation, which are still *terra incognita* to a great extent. The database, in fact, shall allow us to make the most of the existing evidence, to overcome the limits set by traditional print corpora and to push combinatory linguistic and palæographical research a step further. It is our hope that SigLA will make a significant contribution to the field by proving a useful open-access interactive tool allowing researchers more accurately to look into the palæographical, epigraphic, pinacological, as well as linguistic features of the Linear A writing system of Bronze Age Crete. Progress and advancement in these areas will be a major achievement for the study of the Linear A script, the Minoan language, and the cultural backdrop within which the civilisation inhabiting Crete and the Aegean islands flourished in the Late Bronze Age.

References

Bennet, John (2008). "Now You See It; Now You Don't! The disappearance of the Linear A script on Crete." In: *The Disappearance of Writing Systems: Perspectives on Literacy and Communication*. Ed. by John Baines, John Bennet, and Stephen Houston. Sheffield: Equinox Publishing, pp. 1–29.

Chadwick, John (1967). *The Decipherment of Linear B*. Cambridge: Cambridge University Press.

Davis, Brent (2013). "Syntax in Linear A: The Word-Order of the 'Libation Formula'." In: *Kadmos* 52.1, pp. 35–52.

——————— (2014). *Minoan Stone Vessels with Linear A Inscriptions*. Leuven: Peeters.

Decorte, Roeland P.-J. E. (2018). "The Origins of Bronze Age Aegean Writing: Linear A, Cretan Hieroglyphic and A New Proposed Pathway of Script Formation." In: *Paths Into Script Formation in the Ancient Mediterranean*. Ed. by Silvia Ferrara and Miguel Valério. Studi Micenei ed Egeo-Anatolici, Nuova Serie, Supplemento 1, pp. 13–50.

Del Freo, Maurizio and Julien Zurbrach (2011). "La préparation d'un supplément au *Recueil des inscriptions en linéaire A*. Observations à partir d'un travail en cours." In: *Bulletin de correspondance hellénique* 1.135, pp. 73–97.

Godart, Louis and Jean-Pierre Olivier (1976–1985). *Recueil des inscriptions en linéaire A*. Vol. I-V. Paris: Librairie Orientaliste Paul Geuthner.

Hallager, Erik (1996). *The Minoan Roundel and other Sealed Documents in the Neopalatial Linear A Administration*. Vol. 14. Aegaeum. Liège/Austin: Université de Liège/University of Texas.

——————— (2011). "On the origin of Linear B administration." In: *Πεπραγμένα Ι′ Διεθνοῦς Κρητολογικοῦ Συνεδρίου 2006 [Proceedings of the 10th Cretological Conference 2006]*. Ed. by Maria Andreadaki-Vlasaki [Μαρία Ἀνδρεαδάκη-Βλαζάκη] and Eleni Papadopoulou [Ἑλένη Παπαδοπούλου]. Vol. Α1. Χανιά [Chania]: Φιλολογικὸς Σύλλογος «Ὁ Χρυσόστομος» [Literary Society "Chryssostomos"], pp. 317–239.

Judson, Anna (2017). "The decipherment: people, process, challenges." In: *Codebreakers & Groundbreakers*. Ed. by Anastasia Christophilopoulou, Yannis Galanakis, and James Grime. Catalogue of exhibition held Oct. 2017–Feb. 2018. Cambridge: Charlesworth Press, pp. 15–29.

Karnava, Artemis (2016). "On Sacred Vocabulary and Religious Dedications: The Minoan 'Libation Formula'." In: *Metaphysis: Ritual, Myth and Symbolism in the Aegean Bronze Age. Proceedings of the 15th International Aegean Conference, Vienna, Institute for Oriental and European Archaeology, Aegean and Anatolia Department, Austrian Academy of Sciences and Institute of Classical Archaeology, University of Vienna, 22-25 April 2014*. Ed. by Eva Alram-Stern et al. Vol. 39. Aegaem. Leuven-Liège: Peeters, pp. 345–355.

Militello, Pietro (1989). "Gli scribi di Haghia Triada." In: *La Parola del Passato* 44 (2), pp. 126–147.

Pope, Maurice (2008). "The Decipherment of Linear B." In: *A Companion to Linear B: Mycenaean Greek Texts and their World*. Ed. by Yves Duhoux and Anna Murpurgo Davies. Vol. 1. Leuven: Peeters, pp. 1–23.

Raison, Jacques and Maurice Pope (1971). *Index du Linéaire A*. Vol. 41. Incunabula Graeca. Roma: Edizioni dell'Ateneo.

Salgarella, Ester (2019). "Drawing lines: The palaeography of Linear A and Linear B." In: *Kadmos* 58.1–2, pp. 61–92.

——————— (2020). *Aegean Linear Scripts: Rethinking the relationship between Linear A and Linear B*. Cambridge Classical Studies. Cambridge: Cambridge University Press.

Salgarella, Ester (in preparation). "Mix and Match: A combinatory (re-)classification of Linear A signs."

Schoep, Ilse (2002). *The Administration of Neopalatial Crete. A Critical Assessment of the Linear A Tablets and their Role in the Administrative Process*. Salamanca: Ediciones Universidad de Salamanca.

Steele, Philippa M. and Torsten Meissner (2017). "From Linear B to Linear A: The problem of the backward projection of sound values." In: *Understanding relations between scripts: the Aegean writing systems*. Ed. by Philippa M. Steele. Oxford/Philadelphia: Oxbow Books, pp. 93–110.

Thompson, Rupert (2010). "In defence of ideograms." In: *Études mycéniennes 2010: Actes du XIIIᵉ colloque international sur les textes égéens*. Ed. by Pierre Carlier et al. Vol. 10. Biblioteca di Pasiphae. Pisa and Roma: Fabrizio Serra, pp. 545–561.

Tomas, Helena (2006). "Comparing Linear A and Linear B Administrative Systems: The Case of the Roundel and the Elongated Tablet." In: *Colloquium Romanum: atti del XII colloquio internazionale di micenologia*. Ed. by A. Sacconi et al. Vol. 2. Biblioteca di Pasiphae, pp. 767–774.

————— (2010a). "Cretan Hieroglyphic and Linear A." In: *The Oxford Handbook of the Bronze Age Aegean (ca. 3000–1000 BC)*. Ed. by Eric H. Cline. Oxford: Oxford University Press, pp. 340–355.

————— (2010b). "The story of the Aegean tablet: Cretan Hieroglyphic, Linear A, Linear B." In: *Bulletin of the Institute of Classical Studies* 53 (2), pp. 133–134.

————— (2011a). "Linear A Scribes and Their Writings Styles." In: *Pasiphae: rivista di filologia e antichità egee* 5, pp. 35–58.

————— (2011b). "Linear A tablet ≠ Linear B tablet." In: *Πεπραγμένα Ιʹ Διεθνοῦς Κρητολογικοῦ Συνεδρίου 2006 [Proceedings of the 10th Cretological Conference 2006]*. Ed. by Maria Andreadaki-Vlasaki [Μαρία Ἀνδρεαδάκη-Βλαζάκη] and Eleni Papadopoulou [Ἑλένη Παπαδοπούλου]. Vol. A1. Χανιά [Chania]: Φιλολογικὸς Σύλλογος «Ὁ Χρυσόστομος» [Literary Society "Chryssostomos"], pp. 331–343.

————— (2017a). "From Minoan to Mycenaean elongated tablets: defining the shape of Aegean tablets, Aegean Scripts." In: *Proceedings of the 14th International Colloquium on Mycenaean Studies, Copenhagen, 2–5 September 2015*. Ed. by Marie-Louise Nosch and Hedvig Landenius Enegren. Vol. 1. Roma: Istituto di Studi sul Mediterraneo Antico, pp. 115–126.

————— (2017b). "Linear B script and Linear B administrative system—different patterns in their development." In: *Understanding relations between scripts: the Aegean writing systems*. Ed. by Philippa M. Steele. Oxford/Philadelphia: Oxbow Books, pp. 57–68.

Younger, John (2000). "Linear A Texts and Inscriptions in Phonetic Transcription." URL: http://people.ku.edu/~jyounger/LinearA/.

Digitising Swahili in Arabic Script With *Andika!*

Kevin Donnelly

Abstract. In order for traditional culture, as reflected in manuscripts, to make the transition to the digital age, there is a need to use modern technology to make them available. This means more than simply making scans of the manuscripts—it means storing the manuscripts in a digital format which will allow them to be searched, to have concordances and frequency lists compiled, and so on. Where the script used for the traditional material is the same as the current script, this may present few difficulties, but in cases where the traditional material uses a script that is no longer used for the language, this may present difficulties. This paper presents free (GPL3) tools to address these issues for the Swahili language of East Africa (though the general principles are applicable elsewhere), so that heritage material written in a displaced (Arabic) script (S1) can be easily converted to digital form and automatically transliterated to the contemporary (Roman) script (S2).

1. Introduction

This paper addresses ways in which cultural material in a displaced script can be transitioned to the modern digital age. The paper is in three parts.

- The reasons why digitising the *actual text* (as opposed to providing only scans or transcriptions) is essential.
- Tools to do this for Swahili in Arabic script.
- The multitude of ways in which manuscript poetry digitised in this way can be presented.

Kevin Donnelly　iD 0000-0002-0871-6180
Independent reseracher, 7 Ty'n Cae, Llanfairpwll, Ynys Môn, Wales, LL61 6UX, UK
`kevin@dotmon.com`

2. Why Do We Need "Full" Digitisation?

2.1. Script Displacement

The loss of cultural capital due to language displacement is now well-recognised,[1] but a similar loss is caused by script displacement.

In many parts of the world, the scripts formerly used to write particular languages have been superseded by other scripts. This is especially the case for African languages such as Swahili, where over the last century the Roman script has displaced the Arabic script formerly used by literate individuals on the East African coast. Adapting the well-established usage of L1 and L2 to denote "first language" and "second language," we might refer to Swahili in Arabic script as S1, and to Swahili in Roman script as S2.[2]

When historical scripts (S1) are displaced by newer scripts (S2), either as a result of past colonial policies or more recent national policies enforcing orthographic change, a phenomenon of progressive "S1 deliteracy" may occur. This can be defined as a situation where modern-day speakers (especially younger ones) are increasingly unable to read documents that may encode significant amounts of cultural heritage. A wealth of traditional linguistic and cultural material (e.g., poetry, histories, religious tracts) may therefore become increasingly inaccessible to speakers of that language.

Script displacement receives less attention than language displacement, perhaps because it is assumed that S1 cultural material can be conserved via digital scanning of the manuscript, or by creating a digital transliteration (more or less phonetic as the case may be) into S2. However, there are issues with both of these.

2.2. Digital Scans

Digital scans are just so many pictures—they cannot be searched unless they are transcribed. You can change the resolution of the scan on-screen, you can move the page around, you can leaf through the document, but that's about it. They are a great resource for librarians and archivists, in that they allow easier access to manuscripts considered as objects, but they have limited value to scholars of history, language or literature who may be more interested in the content, because they lack the scope for unpacking that content rather than simply looking

---

1. `eldp.net`, `endangeredlanguages.com`
2. This usage could of course be extended if the language involved has ever been written in more than two scripts.

at it. Moreover, their large size makes them difficult to transfer, especially where internet access is limited, and frustratingly slow to navigate through, particularly on older computers.

These problems can to some extent be resolved by converting S1 scans to pdf and enriching them with additional text layers, as Thilo Schadeberg and Ridder Samsom have done for Sacleux (1939). However, selection of text on such pdfs can be haphazard. Moreover, if you add a text layer, which transcription does it use? Standard (modern), or that used in the manuscript? Or both (one text layer for each)? It is also difficult to do any sort of computer-based analysis (eg list all words occurring at the end of a line of poetry), unless you work solely on the text layer. Arguments for creating a text layer are in effect arguments for a stand-alone digitisation.

Another option with digital scans is to create an interface to them that allows annotations to be made on the image, but this raises questions about how should such annotations should be stored, whether they should be an adjunct to the scan or somehow integrated with it, and how they might be searched and compared.

Where S1 has been maintained, another option is to provide an S1 digital version of the text alongside the S1 scan.[3] Sometimes the scan is omitted in favour of a close (diplomatic) S1 transcription of the manuscript, with the interface allowing round-tripping between the transcription and the manuscript.[4] It is at this point that the text leaps, as it were, off the page and into the computer, out of the past and into the present or future—we have the potential to handle the text in the way we handle a modern computer-generated document, but it is still grounded in the original manuscript.

2.3. Transcription Only

Close S1 transcription for S1 originals requires the conventions used in the transfer from page to screen to be defined in detail. But where we have an S1 original and an S2 transliteration, this is even more important. This goes beyond the transliterations of individual letters (e.g., to transliterate the Arabic letter *khah* خ and *shin* ش, German scholars prefer *ḫ* and *š* respectively, while English scholars prefer *kh* and *sh*). More substantively, it raises questions such as:

– How much silent emendation of the text has been done?
– Have sections of the text been omitted, and why?
– Have ambiguous readings been flagged, or simply ignored?

---

3. `ctext.org, beowulf.uky.edu`
4. `rhyddiaithganoloesol.caerdydd.ac.uk, chaucermss.org`

As a thought-experiment, consider whether any linguist, literary scholar, or historian would seriously suggest studying Chinese, Greek, Arabic, Egyptian hieroglyphic texts solely via transliteration. S2 transliteration involves decisions that make the transcriber perforce an editor whose decisions the reader must take on trust. In the case of Swahili, we have in the past often ended up with what looks like an overly "tidy" text, with all lines exactly fitting the metre, all rhymes perfect, and so on. A transliteration-only approach not only wrenches the contents from the context in which they were written, it devalues S1 further, and it balkanises the material (it is scattered over various publications, may use a variety of transliterations, may reflect more or less standardisation, and so on).

It might be argued that combining both of the above methods, by presenting an S2 transliteration alongside an S1 scan, is a viable solution to the shortcomings identified. But this solution is only a partial remedy, because it decouples the medium from the message, losing part of what makes the material a cultural resource. The S1 scan now stands apart from the S2 text, and needs to be periodically reintegrated with it for reading purposes.

In fact, however valuable, all these options (S1 scan alone, S2 transliteration alone, S1 scan + S2 transliteration) tend to suggest that S1 belongs to the past, and has little to contribute to the modern culture. Moreover, a judgement is being made on the "value" of the language, such that peripheral languages (minority languages either in terms of the number of speakers or the political "heft" of those speakers) get downgraded. The implication is that some languages do not "deserve" the resources available to others. As noted above, how many scholars would consider studying Chinese or Arabic solely in Roman transliteration?

2.4.   Full Digitisation

In the past, scans were expensive and impractical. Transcription, with all its shortcomings and value judgements, was therefore seen as the only viable option, even if it was "lossy" when compared with the original manuscript. This is no longer the case: most mobile phones can take high-quality photos, and the ongoing expansion of the Unicode encoding standard[5] makes it possible for virtually any script to be represented by modern computers. There are therefore few reasons nowadays for not producing "full" digitisations (where the text can be fully processed by a computer to allow searches, the creation of wordlists and so on), backing them up where possible with photographs of the manuscript.

---

5. `home.unicode.org`

This cultural material can then transition fully to the modern, digital world, instead of being viewed as an "object" in a museum collection.

The next section of the paper looks at tools which enable this for S1 Swahili (Swahili in Arabic script). The tools are called *Andika!*, the Swahili word for "write!," which often occurs at the beginning of a poem as a command to the scribe to take up his pen and write down the words of the poem.[6] The general principles behind the toolset can be applied to any language where script displacement is an issue.

3. A Toolset to Digitise S1 Swahili

3.1. Swahili

Swahili is possibly the most widely-spoken Bantu language, in terms of both geographical area and number of speakers. It is widely used as L2 by some 90m people in Kenya, Tanzania, Uganda and the DRC, but it is spoken as L1 by perhaps only 2m people (Hinnebusch, 2003) on the East African coast, from Brava in Somalia down to the Comoro Islands off the coast of Mozambique.

The location of the Swahili meant that they became part of the Indian Ocean trading networks from an early period, and in turn this led to their becoming Islamised. The spread of literacy based on the Arabic script led to the writing of their own language in that script, and Swahili has the longest written heritage of any sub-Saharan African language—poetry survives from the late 1600s onwards (Knappert, 1967; 1972; 1982). The greatest flowering of "classical" Swahili literature was in the 1800s, when poets played a role in many of the "city-states" along the coast (Lamu, Pate, Mombasa, Zanzibar, etc). In the late 1800s European missionaries produced Christian material in Arabic script, but under the British colonial administration the language was "standardised" in a Roman orthography from the 1930s on, and since then the use of Arabic script has declined drastically. That does not mean, however, that S1 Swahili has disappeared—it is still used extensively in religious contexts (e.g., mosque schools), and in particular areas. For instance, Ottenheimer (2012, p. 2) notes that "Arabic script is widely used for Shinzwani [a Swahili dialect in the Comoro Islands], with a literacy rate over 90%".

A wide variety of Swahili poetry in different metres has been published, from religious meditations to ballads to love-songs, but there is also a body of prose work that has been less frequently published. Much

---

6. The tools are available under a free (GPL3) license at `kevindonnelly.org.uk/swahili`. The site also includes a manual, and a converter to round-trip between S1 and S2 Swahili—see 3.5 below.

of this S1 material exists in manuscripts, either originals or copies of originals, in Western libraries, and this is probably only a fraction of the extant total—manuscripts are handed down through the generations as family heirlooms.

At present, the only viable way of preserving these S1 (Arabic script) manuscripts is to scan them, or to transcribe them into S2 (Roman script), because the tools available to handle S1 Swahili are limited. Although a word-processor can be set up to use Arabic script, most Arabic fonts do not contain all the glyphs (e.g., *p* /p/, *ng'* /ŋ/) necessary to write Swahili.[7] An additional factor is that the standard Arabic keyboard has a different layout from the standard English (US or UK) keyboard, so using an English keyboard to type Arabic, or vice versa, means trying to mentally translate between the two layouts.

Modern computing platforms give us a viable way to address these issues relatively easily, so that we can type S1 Swahili directly into a computer.

3.2.   Characters for Swahili Sounds

The Unicode Consortium, formed in 1991, has the goal of "support[ing] the writing systems used by all the world's languages [by] provid[ing] a unique code for every character, in every language, in every program, on every platform."[8] As of March 2020, the Unicode Standard encompasses 154 scripts and over 143,000 characters,[9] meaning that a great many characters from Arabic-based scripts are covered.[10]   However, even if a character has been recognised in Unicode, it does not follow that computer fonts will contain that character, and this is the case for most Arabic fonts, which do not contain all the characters necessary to write Swahili.[11]   The most commonly missing sounds, with the *Andika!* character, for them are set out in Table 1.[12]

---

7.   In earlier times, scribes dealt with this deficiency either by using a character that represented a similar sound, or borrowing a character from another Arabic-script language that had a similar sound. So /p/ might be represented by Arabic ب /b/ or Persian پ /p/.

8.   `home.unicode.org/basic-info/overview`

9.   `unicode.org/versions/\index{Unicode}Unicode13.0.0`

10.   `unicode.org/charts/PDF/U0600.pdf`

11.   ISESCO (Islamic Educational, Scientific and Cultural Organization) has proposed a standard Arabic script that would cater for all African languages (Chtatou, 2010), but this tends to ignore local writing traditions (Warren-Rothlin, 2014)—for example, the proposed vowel for *e* seems to be used only in Fulfulde.

12.   The last three are for representation of northern dialects.

TABLE 1. Swahili characters missing from most Arabic fonts

p	ch	g	ng'	v	t^r	d^r	zh
پ	خ	غ	نغ	ڤ	ٹ	ڈ	ژ

In such a case, there are then two options. One is to add that character to the desired font using a font editor.[13] But the simpler option is to use a comprehensive Arabic font that contains that character. *Andika!* uses SIL's Scheherazade font.[14] One important point is that since S1 Swahili is usually vocalised, it is best to avoid fonts which use Arabic ligations extensively, since these can cause problems with placement of the vowel signs. Even a font like Scheherazade, though, is still missing characters used by some scribes (e.g., *noon with teh above*, as used in Chimwiini in the most northerly part of the Swahili littoral). In that case, the most workable option in the short term is to add the character by hand using a font editor, and seek in the longer term to have the codepoint added to the Unicode Standard, and then to the font.

3.3.   Accessing the Characters

Having chosen a font which contains all the characters needed to represent Swahili, the next requirement is a way to access those characters via the computer keyboard. Since the standard Arabic keyboard has a different layout from the standard English (US or UK) keyboard, switching between them means memorising different keys for the same sounds for each script. For example, *t* is on the top row under the left hand on an English keyboard, but *teh* ت is in the middle row under the right hand on an Arabic keyboard.

To avoid this, *Andika!* uses a key layout for the Arabic characters (Figure 1) that maps to the layout of the English keyboard, meaning that typists can leverage what they already know from typing S2 standard Swahili. The characters are grouped as logically as possible, using either sound or character likeness. For instance, *sukun* is on the full stop key, and short vowels, long vowels, and vowel carriers are all on the same key. Related Arabic characters are mostly on the same keys as the English characters. For instance (Figure 2), *dal* د is on the *D* key, *dhal* ذ is accessed using *Shift+D*, and *dad* ض using *AltGr+D*. A character repre-

---

13.   designwithfontforge.com/en-US/Adding_Glyphs_to_an_\index{Arabic}Arabic_Font.html If there is not already a Unicode codepoint for that character, a codepoint in a Private Use Area can be used: en.wikipedia.org/wiki/Private_Use_Areas.

14.   software.sil.org/scheherazade. Other possibilities are Khaled Hosny's Amiri font and the PakType fonts.

FIGURE 1. The Swahili keyboard layout in *Andika!*

senting the alveolar *d* as used in Mombasa, *dal with tah above* ڐ, borrowed from Urdu, can be accessed using *AltGr+Shift+D*.

FIGURE 2. Character cluster on the *D* key

The result is that S1 (Arabic) Swahili can be typed as quickly and easily as S2 (Roman) Swahili, and on the same keyboard. The same approach could be used for any language to map S1 characters to the relevant S2 keyboard.

3.4. Using S1 Swahili

Now that we have a means of representing Swahili in Arabic script, two forms of use are possible: using it to write contemporary Swahili, and using it to replicate historical Swahili manuscripts in a digital format. The remainder of this section discusses the first use-case, and the next section discusses the second.

Before that, it may be worth emphasising that Arabic script is just as capable as Roman script at representing any language. There are no purely linguistic or graphemic reasons for favouring Roman script over Arabic script for the representation of Swahili: the fact that Swahili is overwhelmingly written in Roman script is due purely to political and historical developments, and not to any shortcoming in the Arabic script. As the British Library's Endangered Archives Programme

says, "Ajami, the modified Arabic scripts used in writing African languages, have been deeply embedded in the history and culture of many Islamized societies of Africa,"[15] and Mumin (2014, p. 44) notes that the use of Arabic script has been attested for at least 80 African languages. As with Roman, Cyrillic and other scripts, Arabic script has added additional characters or diacritics when necessary to cater for languages as different as Persian, Turkish, Kurdish, Pashto, Urdu, Malay, Hausa, etc., as Warren-Rothlin (2014, 269ff) points out. In light of this, comments about "the incongruence between Swahili and Arabic and the resulting incompatibility of the Arabic script to write Swahili" (Vierke, 2014, p. 326) are misleading, and indeed seem to be a manifestation of the point made in the introduction about S1 being seen as belonging to the past.

The script itself is not the problem. Rather, its usage has been hampered by a lack of standardisation, where writing conventions tended to be ad hoc responses to recording speech (for a similar issue in West Africa, see Warren-Rothlin (2014, 269ff)). S2 is more likely to have such conventions than S1, given that in many cases S2 may have been expressly designed to handle the language, perhaps via a committee, whereas S1 is more likely to have been progressively developed and adapted over a period of time by individuals unable to do anything but promote good practice as they see it.

This issue of standardised spelling is relevant when discussing contemporary use of S1 Swahili. Although S2 Swahili is now the standard used by millions of speakers, and that will not change, the ability to use S1 may be useful in domains (e.g., mosque schools) where that script is still used, or in places (e.g., the Comoros) where S1 literacy is still high. The key point is to allow the *option* of using S1 for cultural heritage purposes. A key practical issue for contemporary use, however, is avoiding the additional work involved in typing the text twice, once in each script. Preferably, it should be possible to get either S1 or S2 "for free" from the other, meaning that the same text can be created in either script, and converted to the other as required This also provides an easy way of increasing the amount of modern S1 text available, for instance, by converting S2 webpages or other documents to S1. Crucially, however, conversion is only possible if both scripts use standardised spelling.

Since there is currently no standard for S1 Swahili spelling, *Andika!* uses the proposed system set out in Omar and Frankl (1997).[16] Combined with the keyboard layout described above, this means that text

---

15. eap.bl.uk/project/EAP1042

16. Some slight modifications have been made. For instance, the authors suggest rules for omitting short vowels, but it is actually quicker to type them, and this also simplifies conversion.

can be typed into a word-processor at around the same speed in either
script using almost exactly the same keys (Figure 3)—the only difference
is the need to type a capital letter to get a long vowel in the penultimate
syllable in S1.

S2 standard:	ninakwenda nyumbani sasa
S1 typing:	ninakweEnda nyumbaAni saAsa
S1:	نِنَكوِيندَ نيُمبَانِ سَاسَ
English:	I am going home now

FIGURE 3. Typing S1 Swahili

3.5.   Converting Between S1 and S2

In the S1 → S2 direction (Figure 4), S1 is converted first to a Romanised
abstraction, and then converted to a standard S2 transliteration.

The reason for the intermediate abstraction is to offer scope for mul-
tiple transliterations. For instance, when dealing with older manu-
scripts (see 4.2 below) we may wish to have a close transliteration of the
Arabic script as well as a standard transliteration. Alternatively, it may
be appropriate to replace a standard transliteration with one that reflects
dialectal features. For instance, if a scribe has written ذِيچَ, dhīcha, the
equivalent in the northern Bajuni dialect to standard Swahili *vita*, 'war',
a transliteration such as *ẕitʲa* might be preferred, in order to come as
close as possible to standard S1 while giving an indication of the dialec-
tal pronunciation.[17] Another option would be to add a transliteration
for Arabic, to handle bilingual Arabic/Swahili text (e.g., *Qasida Hamziyya*
poems—see 4.2 below). Currently, Arabic text is transliterated using the
close transliteration conventions for Swahili, and some features of the
Arabic language are not optimally handled in this.

In the S2 → S1 direction, no intermediate abstraction is currently
used, because there is only output at present—the proposed standard
spelling in Omar and Frankl (1997). Nevertheless, the same approach
could be used, so that different S1 spellings are supported.

Virtually no editing is required in either direction, with the exception
that in S1 → S2 capital letters need to be added where appropriate, since
Arabic has no concept of capital letters. The website[18] gives an example
of a browser-based frontend to the converter code, where text can be
copied and pasted into a box, and converted to the other script.

17. *zi-* is a northern variant of the standard class 8 marker *vi-*.

18. `kevindonnelly.org.uk/swahili`
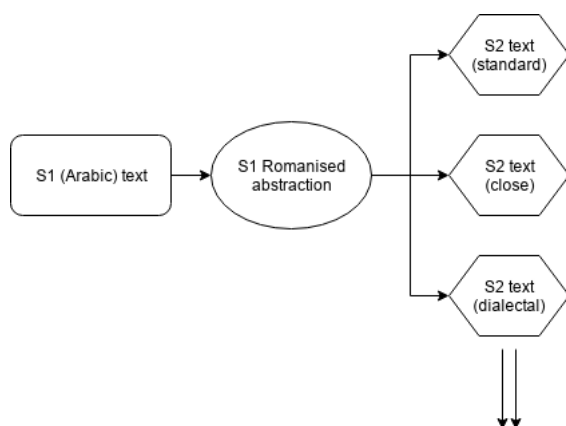
FIGURE 4. Conversion from S1 to S2

4.   Digitising Heritage Manuscripts

With tools in place to type S1 Swahili into a computer, and to convert it to a variety of S2 Swahili transliterations, we now have a means of digitising S1 manuscripts, particularly those containing traditional poetry. This section demonstrates replicating the content of the manuscript in digital form, and gives examples of various types of enriched output.

The digitisation of heritage manuscripts has additional requirements compared to the digitisation of contemporary texts. For instance, we may want to reflect the layout on the physical page, or add alternate readings, or draw attention to scribal errors, or add contextual or etymological notes on individual words, or add a translation. Likewise, we are almost certain to need a list of the words in the manuscripts, so that concordances, indexes or other editorial matter can be prepared. The solution proposed by *Andika!* is to insert each word of the manuscript text into a database, so that additional material like this can be added at word-level. Subsets of the material can then be retrieved from the database as required, in whatever format is appropriate.

4.1.   The Digitisation Process

A key difference between traditional transliteration and the *Andika!* approach is that the process begins with typing out the manuscript itself instead of typing out a transliteration of the manuscript. The latter step is not required, because the transliteration (indeed, several transliterations—see 3.5) can be generated automatically from the re-typed manuscripts. The process is summarised in Figure 5.

FIGURE 5. Digitisation process for S1 manuscripts

In Step 1, the manuscript is typed out as if it were a piece of contemporary text, but instead of trying to follow a spelling standard, we type only what the scribe wrote in the manuscript. To simplify import into the database, the typed text should follow a specific format—each rhymed stretch of the poem should appear on a line by itself, blank lines should be inserted after stanzas, etc. Figure 6a shows stanzas 16 and 17 from an original manuscript version of the Swahili *Ballad of Jaʾfar*. Figure 6b shows the same stanzas typed out and ready for import into the database.

FIGURE 6. (a) Manuscript stanzas from the Ballad of Jaʾfar. (b) The stanzas as typed

In Step 2, *Andika!* parses the typed copy of the manuscript, and imports the lines of the poem into a database table. In the process both a standard and a close S2 transcription for the S1 text is automatically generated (Figure 7). Then each line is split into words, which are imported into another database table (Figure 8a).

In Step 3, each word in the database can be inspected to correct the automatic transcription if necessary. Individual words can then be an-

notated with notes, alternate readings, corrected transliterations, etc., and a translation can be added, as in Figure 8a, keyed to the first word of the line. This allows the development of a full critical apparatus for the text.

In Step 4, the textual material in the database can be exported in PDF format, with various options for text layout, text colouring, line numbering, translation position, etc. Figure 8b shows the two stanzas output as single lines, with S1 Swahili in green, stanza numbering, an English translation, and a generated S2 Swahili transcription. A generated close transcription (in green) is also included, output right-to-left by word so that it matches directly with the S1 script above it.

msno	stanza	loc	arabic	close	standard
16	16	a	هُرُم زكَمْشِكَ	huruma zikamshika	huruma zikamshika
16	16	b	عَلِيْ كَشَوِشِكَ	ʻalii kashawishika	alii kashawishika
16	16	c	كَمْبَ فَتُمَ پُلِكَ	kamba fatuma pulika	kamba fatuma pulika
16	16	d	اُلِلنَلَ نَمْبِيَ	ulilonalo nambiya	ulilonalo nambiya
17	17	a	ان كِثْو هُكُوْم	una kitwa hukuuma	una kitwa hukuuma
17	17	b	أَمْ أمَшِكْوَ ن خُمَ	ama umeshikwa ni ḥuma	ama umeshikwa ni huma
17	17	c	اكَمْجِبُ فَتُمَ	akamjibu fatuma	akamjibu fatuma
17	17	d	كْو مَرَض سِكُلِيَ	kwa maraḍi sikuliya	kwa maradhi sikuliya

Figure 7. The stanzas from Figure 6 imported as lines

4.2.   Other Examples

Mwana Kupona is one of the few female Swahili poets whose work has come down to us. In 1858 she wrote a poem containing advice for her daughter, and stanza 6 reads (in standard S2 Swahili):

> mwana adamu si kitu, na ulimwengu si wetu,
> walau hakuna mtu ambao atasaliya
>
> *mankind is as nothing, and the world does not belong to us,*
> *and there is no person who will live forever*

Below is an S1 transliteration and an automatically generated close transcription of this stanza from the first of two manuscripts (Figure 9).

مَانَ اَدَامُ سِكِتُ ٭ نَوُلِمِغُ سِوِتُ ٭ وَلَوُ هَكُوْنَ مْتُ ٭ اَبَوُ اَتَسَلِيَ

māna aḏāmu sikiṯu ٭ nawulimiḡu siwiṯu ٭ walawu hakūna mṯu ٭ abawu aṯasaliya

msno	stanza	loc	position	arabic	close	standard	english
16	16	a	1	هُرُمَ	huruma	huruma	Ali was seized with pity,
16	16	a	2	زكَمْشِكَ	zikamshika	zikamshika	
16	16	b	1	عَلِيْ	ʿalii	alii	and became perplexed.
16	16	b	2	كَشَوِشِكَ	kashawishika	kashawishika	
16	16	c	1	كَمْبَ	kamba	kamba	He said: Fatima, listen --
16	16	c	2	فَتُمَ	fatuma	fatuma	
16	16	c	3	پُلِكَ	pulika	pulika	
16	16	d	1	أُلِلْنَلْ	ulilonalo	ulilonalo	tell me what's wrong with you.
16	16	d	2	نَمْبِيَ	nambiya	nambiya	
17	17	a	1	أَنَ	una	una	Do you have a headache,
17	17	a	2	كِتْوَ	kitwa	kitwa	
17	17	a	3	هُكُوْمَ	hukuuma	hukuuma	
17	17	b	1	أَمَ	ama	ama	or have you a temperature?
17	17	b	2	أُمِشِكْوَ	umeshikwa	umeshikwa	
17	17	b	3	ن	ni	ni	
17	17	b	4	حُمَ	ḥuma	huma	
17	17	c	1	أَكَمْجِبُ	akamjibu	akamjibu	And Fatima replied:
17	17	c	2	فَتُمَ	fatuma	fatuma	
17	17	d	1	كُوْ	kwa	kwa	I am not crying because I am ill.
17	17	d	2	مَرَض	maraḍi	maradhi	
17	17	d	3	سِكُلِيَ	sikuliya	sikuliya	

(a)

(۱۷) هُرُمَ زِكَمْشِكَ ٭ عَلِيْ كَشَوِشِكَ ٭ كَمْبَ فَتُمَ پُلِكَ ٭ أُلِلْنَلْ نَمْبِيَ

nambiya ulilonalo ∗ pulika fatuma kamba ∗ kashawishika ʿalii ∗ zikamshika huruma

(**16**) huruma zikamshika ∗ Aliyi kashawishika ∗ kamba Fatuma pulika ∗ ulilo nalo nambiya

Ali was seized with pity, and became perplexed. He said: Fatima, listen—tell me what's wrong with you.

(۱۸) أُنَ كِتْوَ هُكُؤُمَ ٭ أَمَ أُمِشِكْوَ نِ حُمَ ٭ أَكَمْجِبُ فَتُمَ ٭ كْوَ مَرَضٍ سِكُلِيَ

sikuliya maraḍi kwa ∗ fatuma akamjibu ∗ ḥuma ni umeshikwa ama ∗ hukuuma kitwa una

(**17**) una kitwa hukuuma ∗ ama umeshikwa na huma ∗ akamjibu Fatuma ∗ kwa maradhi sikuliya

Do you have a headache, or have you a temperature? And Fatima replied: I am not crying because I am ill.

(b)

FIGURE 8. (a) The stanzas from Figure 6 imported as words. (b) The stanzas from Figure 6 exported as a fully digital text
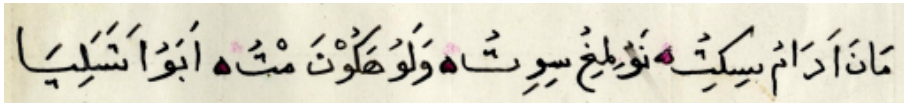
FIGURE 9. Utenzi wa Mwana Kupona, stanza 6, first manuscript.
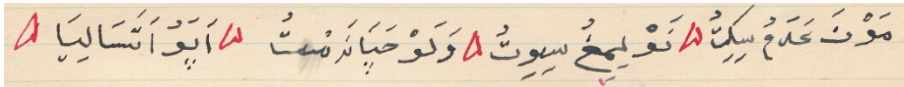
FIGURE 10. Utenzi wa Mwana Kupona, stanza 6, second manuscript.

Figure 10 shows the same stanza from a second manuscript. The S1 transliteration and transcription below show notable spelling differences compared to the first manuscript (e.g., the use of *ʿayn* ع and *ḥa* ح), and a typo in the first word, where *fatha* and *sukun* have been reversed.

مَوْنَ عَدَمُ سِكِتُ ٭ نَوْلِمِغُ سِوِتُ ٭ وَلَوْ حَپَانَ مْتُ ٭ اَپَوْ اَتَسَالِيَا

mawna ʿaḏamu sikiṯu ٭ nawlimiḡu siwiṯu ٭ walaw ḥapāna mṯu ٭ apawu aṯasāliyā

Sayidi Abudallah (Hichens, 1939) wrote a lament in 1853 about the declining fortunes of the coastal city-state of Pate. The two stanzas shown in Figure 11, given here in S2 Swahili, describe the opulence of the town in its better days:

> Nyumba zao mbake zikinawiri kwa taa za kowa na za sufuri;
>   masiku yakele kama nahari, haiba na jaha iwazingiye.
> Wapambiye swini ya kuteuwa, na kulla kikombe kinakishiwa;
>   kati watiziye kuzi za kowa katika mapambo yanawiriye.
>
> Their homes were brightly lit with lamps of mother-of-pearl and copper;
>   the nights stayed bright as day, beauty and privilege surrounded them.
> They decorated their fine porcelain, and every goblet was engraved;
>   in the centre they placed mother-of-pearl carafes, to glitter amongst the fine things.

The transcription and close transliteration below show how the Arabic script is only a partial representation of the sounds of Swahili. Three vowel glyphs are used to represent Swahili's five vowels (e.g., *yakili* for *yakele*, 'stayed', *kuwa* for *kowa*, 'shell'), and prenasalised consonants are not distinguished (*yuba* for *nyumba*, 'house', *mapabu* for *mapambo*, 'decorative objects').
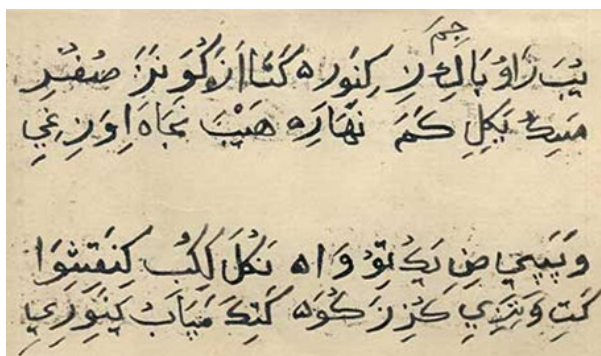
FIGURE 11. Two stanzas from al-Inkishafi.

يُبَ زَاوُ بَاكِ زِكِنَورِ ٭ كَتَاآ زَكُوَ نَزَضُفُرِ

yuba zāwu bāki zikinawiri ٭ katāa zakuwa nazaḍufuri

مَسِكُ يَكِلِ كَمَ نَهَارِ ٭ هَيْبَ نَجَاهَ اِوَزِغِي

masiku yakili kama nahāri ٭ hayba najāha iwazighiyi

وَيَپِي صِنِ يَكُتِوُوَا ٭ نَكُلَ كِكُبِ كِنَقِشِوَا

wapapiyi ṣini yakutiwuwā ٭ nakula kikubi kinaqishiwā

كَتِ وَتِزِي كُزِ زَكُوَ ٭ كَتِكَ مَپَابُ يَنَوِرِي

kati watiziyi kuzi zakuwa ٭ katika mapābu yanawiriyi

Figure 12 shows part of a manuscript by the late Sh. Yahya Ali Omar recording fishing songs in kiBajuni or kiTikuu, a northern Swahili dialect (Donnelly and Omar, 1982).

The automatic S2 transliteration uses a variant of the close trancription, so that it stays as close as possible to standard Swahili orthography, while still representing the distinctive sounds of kiTikuu (see 3.5).

The *Ballad of Mkunumbi* (Harries, 1967) is one of the few books of Swahili poetry to include the S1 text of the manuscript.

This digitisation of the stanza in Figure 13 shows a close transcription only, keyed to the half-line rather than the full line. It also adds stanza numbers in eastern Arabic numerals for the S1 transcription, and in western Arabic numerals for the transcription, where the half-lines are also indicated.
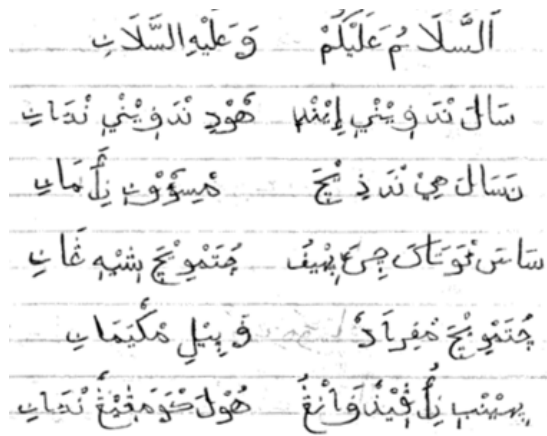
اَلسَّلَا مُ عَلَيْكُمْ ۞ وَ عَلَيْهِ السَّلَانِ

assalāmu ʾalaykum ۞ wa ʾalayhi assalāni

سَالَ نْدَ وِيْنِي إِنْدِ ۞ هُوْدِ نْدَ وِيْنِي نْدَانِ

sāla nḍa wēnye īnḍe ۞ hōḍi nḍa wēnye nḍāni

نَ سَالَ هِيْ نْدَ ذِيْجَ ۞ مْسِوُوْنِ نِأَمَانِ

na sāla hii nḍa ẓīṯ‍ia ۞ msiwōne niamāni

سَاسَ طْوَتَاكَجِئَ پِيْفُ ۞ چُتَمْوِيْجَ شِيْهِ غَانِ

sāsa ṭwaṭākaṯ‍ia pēfu ۞ t‍ʲuṭamwīṯ‍ia shēhe gāni

چُتَمْوِيْجَ مْفِرَاذُ ۞ وَ پِيْلِ مْكَيَمَانِ

t‍ʲuṭamwīṯ‍ia mfiraḍo ۞ wa pīli mkoyamāni

پِيْنْبِ نِ أُقِيْذُ وَانْغُ ۞ هُوْلَ كْوَ مَغِيْغُ نْدَّانِ

pēṃbe ni uw̰ēẓo wāngu ۞ hūla kwa magēgo nḍāni

FIGURE 12. Stanza from a Bajuni fishing song.

١ دُوْلَ مْبِلِ زِلِوَانَ شِكُو نَاسِمْبَ مْبَوَانَ

dōla mbili ziliwāna shikuwe nāsimba mbawāna 1b/a

كَمَتِزُ كُشِنْدَانَ مْتانَ نَلَيْلِيَ

kamaṭezo kushinḍāna mṯāna nalayliya 1d/c
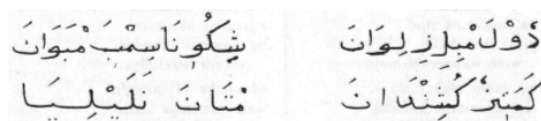
FIGURE 13. A stanza from the Ballad of Mkunumbi

The *Ballad of Rasi ʾlGhuli* was written around 1850 by Mgeni bin Faqihi, and at over 4,500 stanzas is the longest Swahili ballad in existence (van Kessel, 1979). This digitisation of stanza 2,280 (Figure 14) has a close transcription keyed to the half-line, and a standard transcription keyed to the full line.
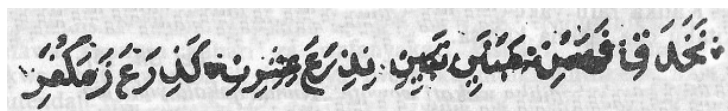
FIGURE 14. Stanza 2,280 from the Ballad of Rasi ʾlGhuli

مَبَنَي تَبَيِن	نَخَدَقٍ فَهَمُنِ	٢٢٨٠
mabanayi tabayini	nakhadaqi fahamuni	2280b/a
na khandaqi fahamuni ∗ mpanaye tabaini		2280a/b

كَذِرَعَ زَمَكُفَر	نِذِرَعَ عِشِرِن	
kadhiraʾa zamakufari	nidhiraʾa ʾishirini	2280d/c
ni dhiraa ishirini ∗ kwa dhiraa za makufari		2280c/d

Qasidas are panegyric poems in Arabic eulogising the Prophet (Knappert, 1971; Sperl and Shackl, 1995). The *Qasida ya Burda* was composed in Arabic by Muhammad bin Saʿidi al-Busiri in the 1300s, and rendered into Swahili verse by Muhammad bin Athumani Hajji al-Hilali Mshela, 1840-1930 (wa Mutiso, 1996). The digitisation of the manuscript in Figure 15 has been set to show the original text in Arabic in blue.

The *Qasida Hamziyya* (so called because it rhymes in *hamza* ء) was composed in Arabic by Muhammad bin Saʿidi al-Busiri (1212-1294), and rendered into Swahili verse by Aidarus bin Athumani bin Ali bin Sheikh Abubakar bin Salim in (probably) 1749 (Knappert, 1968; Parkar, 2020). The digitisation of the manuscript in Figure 15 show the original text in Arabic in blue, and no transcription.

أَمِنْ نَذَكُّرِ جِيْرَانٍ نِذِيْ سَلَمْ ٭ مَزَجْتَ دَمْعَاجَرَى مِنْ مُقْلَةٍ بِدَمْ   ١

نِكَكُكُمْبُكَ جِرَنِ نْيمَ ٭ وَلِيْكُ هَبْ نِذِى سَلَمْ

أُمِلِتَنْغَنْيِ تُرِ كُوْ دَمْ ٭ كُوْمْبَ مَعْنَاي نِهَيْ سِيْمَا

FIGURE 15. First stanza of the Qasida ya Burda

لَمْ يُسَاوُكَ فِيْ عُلَاكَ وَقَدْ حَالَ ٭ سَنَي مِنْكَ دُوْنَهُمْ وَ سَنَاءُ

lam yusāwuka fi ʿulāka waqad ḥāla ٭ sanay minka dūnahum wa sanāʾu

كَوَفَنِ نَوِ رِفْعَانِ بَحَجِزِلِ ٭ نُوْرُ نَرُفْعَةَ كَتِكِنُ كُلُ عَظِيْمَ

kawafani nawi rifʿāni baḥajizili ٭ nūru naruʿfaṭa katikinu kulu ʿaẓīma

They are not equal to you in your elevated status,
the light and sublimity in you is great (in all respect).

FIGURE 16. Second stanza of the Qasida Hamziyya

Figure 16 shows a digitisation of second stanza of the poem. Here again the Arabic original is in blue, and there is an English translation. Both have a close transcription (though the Arabic one, as noted in 3.5, is less than optimal).

Andika! also allows multiple copies of the same poem to be presented in parallel. Below are S1 digitisations of two manuscript versions of the *Ballad of Jaʾfar* (see 4.1 above), each coloured differently, so that they can be compared (in this example, the third line of the stanza differs in each version). Each version has an S2 standard transliteration, keyed to the stanza, and a close transcription (coloured to match the S1 text) keyed

to the S2 word. An English translation is attached to the first (Y) manuscript version.

<div dir="rtl">

(٢٦) كَمْجِبُ كْوَ لِسَنِ ٭ مْطُي سِمْبَئِنِ ٭ پِطْ أُمْپِيْ نَنِ ٭ أُنِيَپْ تَهَرِضِيَ

</div>

tʰariḍiya unipapo ٭ nani umpee peṭe ٭ simbaini mṭuye ٭ lisani kwa kamjibu

Y 25 [23] (**26**) kamjibu kwa lisani ٭ mtuye simbaini *pete umpee nani ٭ unipapo taridhiya
She replied forcefully: I will not disclose that person.
Who have you given the ring to? [Only] when you give [it to me] will I be satisfied.

<div dir="rtl">

كَمْجِبُ كْوَا لِسَنِ ٭ مْتُي سِمْبَاءِنِ ٭ پِتِ يَكُ يَكَنْدَانِ ٭ أُنِيَپْ تَرِظِيْيَا

</div>

tariẓiyā unipapo ٭ yakʲandāni yaku piti ٭ simbaini mtuyi ٭ lisani kwā kamjibu

R 26 [26] kamjibu kwa lisani ٭ mtuye simbaini ٭ pete yako ya chandani ٭ unipapo taridhiya

4.3.   Beyond the Manuscript

Storing manuscript text in a database, as *Andika!* does, opens up some interesting possibilities. As noted earlier, one important side-effect is the easy creation of concordances and indexes. If the wordstores for a number of different manuscripts are combined (in effect, creating a searchable literary corpus), we get a multiplier effect: we are working across manuscripts instead of within them. Such a corpus would, for instance, allow scholars to:

– study character usage and spelling conventions, which may help clarify the genealogy of particular manuscripts;
– trace the occurrence of particular words and examine vocabulary usage in general, which may identify particular schools or authors;
– analyse textual features such as syntactic structure, which could be useful in researching diachronic and synchronic variation;
– consider the usage of fixed expressions (formulae), which may give clues about the process of composition and recital.

As an example, studying the wordstore for the *Ballad of Jaʾfar* referred to several times above allows some significant conclusions to be drawn:

– the verbal consecutive (non-time specific) marker occurs in 30% of all verbforms, reflecting the emphasis on timeless action in the ballad;
– around a quarter of the words in the ballad are derived from Arabic;

- Arabic words are more likely to occur in rhyming positions in stanza-internal lines, suggesting that considerations of rhyme and metre are their main rationale;
- one in five of the verbs used relate to speaking (*say, reply, speak, greet*, etc.);
- almost half of the stanza-internal lines use just three rhymes (*-ni, -ri, -ka*);
- ready-made rhyme-sets seem to be available that will allow the reciter to refer to one of the characters saying something, and bring in a reference to God if appropriate.

5.   Conclusions

This paper has argued that "full" digitisation of S1 heritage manuscripts is the only approach that will liberate the cultural riches locked in them, and avoid them being seen as museum objects that belong in the past, defined solely by type of paper, ink composition, and layout. Such manuscripts are more than just scanned images—they are unique snapshots of a nexus of cultural ideas that still speak to the present and future, even if they were produced in the past. To engage with these ideas, we need to pay these manuscripts the courtesy of transitioning them fully to the digital age, so that we can bring to bear all the tools now available to us in unlocking their meaning.

The *Andika!* toolset described here is one possible way in which this concept could be executed. It is still a work-in-progress, but the concept could be adapted to any language where cultural material is available in a displaced script. Funders of humanities research might also consider the benefits of generating local employment in the "knowledge industry" by paying local people to type heritage manuscripts into a computer alongside the work already being done to scan manuscripts.

References

Chtatou, Mohamed (2010). *Using Arabic script in writing African languages, revisiting ISESCO's experience 25 years later: Field successes and shortcomings*. Paper presented at the workshop "The Arabic Script In Africa: Diffusion, Usage, Diversity And Dynamics Of A Writing System," University Of Cologne.

Donnelly, Kevin and Yahya Ali Omar (1982). "Structure and association in Bajuni fishing songs." In: *Genres, Forms, Meanings: Essays in African Oral Literature*. Ed. by Veronika Görög-Karady. Vol. 1. JASO Occasional Papers, pp. 109–122.

Harries, Lyndon (1967). *Utenzi wa Mkunumbi. A Swahili Potlatch—The Poem about Mkunumbi*. Nairobi: East African Literature Bureau.

Hichens, William (1939). *Al-Inkishafi: The Soul's Awakening*. London: Sheldon Press.

Hinnebusch, Thomas J. (2003). "Swahili." In: *International Encyclopedia of Linguistics*. Ed. by William J. Frawley. Oxford: Oxford University Press, pp. 99–106.

Knappert, Jan (1967). *Traditional Swahili Poetry*. Leiden: Brill.

———— (1968). "The Hamziya deciphered." In: *African Language Studies* 9, pp. 52–81.

———— (1971). *Swahili Islamic Poetry*. Leiden: Brill.

———— (1972). *A Choice of Flowers: Swahili Songs of Love and Passion*. Portsmouth, NH: Heinemann.

———— (1982). *Four Centuries of Swahili Verse: A Literary History and Anthology*. Portsmouth, NH: Heinemann.

Mumin, Meikal (2014). "The Arabic script in Africa: Understudied literacy." In: *The Arabic Script in Africa: Studies in the Use of a Writing System*. Ed. by Meikal Mumin and Kees Versteegh. Leiden: Brill, pp. 41–76.

Omar, Yahya Ali and P. J. L. Frankl (1997). "An historical review of the Arabic rendering of Swahili, together with proposals for the development of a Swahili writing system in Arabic script." In: *Journal of the Royal Asiatic Society*. 3rd ser. 7.1, pp. 55–71.

Ottenheimer, Harriet J. (2012). "Ideology and orthography. Dictionary construction and spelling choice in the Comoro Islands." In: *Études océan Indien* 48. DOI: https://doi.org/10.4000/oceanindien.1521.

Parkar, Ahmed (2020). "Manuscripts and Transmission of Knowledge in Swahili Society: A Comparative Analysis of Form and Usage of Qaṣīda al-Hamziyya." PhD thesis. University of Hamburg.

Sacleux, Charles (1939). *Dictionnaire swahili-français*. Vol. 36. Travaux et mémoires de l'Institut d'Ethnologie. Paris: Institut d'Ethnologie.

Sperl, Stefan and Christopher Shackl, eds. (1995). *Qasida Poetry in Islamic Asia and Africa*. Leiden: Brill.

van Kessel, Leo (1979). *Utenzi wa Rasi ʾlGhuli*. Dar-es-Salaam: Tanzania Publishing House.

Vierke, Clarissa (2014). "Akhi patia kalamu: Writing Swahili poetry in Arabic script." In: *The Arabic Script in Africa: Studies in the Use of a Writing System*. Leiden: Brill, pp. 319–339.

wa Mutiso, Kineene (1996). "Archetypal Motifs in Swahili Islamic Poetry: Kasida ya Burudai." PhD thesis. University of Nairobi.

Warren-Rothlin, Andy (2014). "West African scripts and Arabic-script orthographies in socio-political context." In: *The Arabic Script in Africa: Studies in the Use of a Writing System*. Leiden: Brill, pp. 261–289.

A Semantic Index
for a Dongba Script Database

Duoduo Xu

Abstract. Dongba script is a pictographic writing system still in use in South-West China. The Dongba dictionaries provide valuable resources and should be used in complete synergy to build up a character repertoire. A semantic index with the pictographic radicals, along with other kinds of indexes, is necessary for an ideographic writing system such as Dongba script, while the studies on Dongba radicals are at a preliminary stage. This semantic index also contributes to show that similar markers arose in a logographic system.

1. Introduction

The Dongba script has been enlisted in the World Memory Heritage by the Unesco in 2003. It is a unique pictographic writing system, still in use today.[1] Its pictographs are used to convey in their own way, basically through semantic-visual stylized signs and much less through phonograms, the Naxi thought and language (a member of the Yi, aka "Loloish," branch of the Tibetan-Burman language family), see Ramsey (1987, pp. 249–250). With the recognition of this writing system, many other pictographic scripts have been brought into light. Among other places, South-West China shows to be a cradle of pictographic writings, e.g., Ersu Shaba script (Sun, 1982), Pumi Hangul script (Song, 2010), Muya Sujowu script (Liu and Huang, 2013), Namuzi Pabi/Pazi script (Wang and Wang, 2013), Lhoba Niubu script (Zhao, 2014), etc.[2] Scholars pointed out that the development of early writing systems can be reconstructed through the analysis of these scripts belonging to ethnic groups (Song, 2010).

Many "scribal models" around the world share similar techniques in conveying information. For example, the so-called Dakota winter counts [Lakota Sioux lineage groups, U.S.] (Howard), Mi'q-m'aq

---

Duoduo Xu   ⓘD 0000-0001-6734-8405
Nanyang Technological University
E-mail: duoduo.xu@ntu.edu.sg

1. Revised Proposal for Encoding Naxi Dongba Pictographic Script in the SMP of the UCS, 2011.
2. More scripts can be found in Zhao (2013).

(= Micmac) [Atlantic seaboard / Quebec North-shore, Canada], and Rongorongo [Rapa Nui / Easter Island; modern Chilean territory], seem inventories of certain dates. Kuna (= Cuna) Mola [San Blas Islands on the Atlantic side of Panamá, with minor settlements in the Chucunaque region of the Darien forest, near the Colombian border], used to be body painting patterns.

These scribal symbols show different stages of the development of writing systems. While most pictographic scripts are strictly attested in specific documents, such as indigenous religious manuscripts, calendars, or songs (e.g., Poya Script of Zhuang People in Guangxi Province, China), Dongba script is used also to transcribe local language besides Dongba scriptures. Moreover, Dongba script possesses a greater number of glyphs, if compared to the others.

Moso/Na People 'live' at the border between orality and literacy.[3] The correspondence between the Dongba glyphs and the Moso/Na language they transcribe is different from the one between well-known mature writing systems and their related languages, such as Chinese, Japanese, Latin, Sanskrit, Arabic, etc. Traditionally, Dongba glyphs are used for scriptures. In these works, the pictograms transcribe only the keywords of a verse in Dongba chants (Fu, 1982, p. 6; Wang, 1988, p. 124; Yu, 2009, p. 19). Dongba priests can read out the complete verses from the scriptures, according to the oral versions of the texts learnt by heart (Li, 1997, pp. 70–71).[4]

Dongba writings has also undergone 'evolution'. Dongba glyphs have been used to transcribe secular documents (notation of accounts, land contracts, letters, etc.) in which they are utilized as phonetic loans to write down each syllable of a sentence (Yu, 2003, pp. 252–282; Yu, 2008, pp. 124–250). In the contemporary world, Dongba pictographs are applied to transcribe couplets, modern vernacular language, and even for-

---

3. "Moso" is the old appellation of this ethnic group living along the Jinsha River, the main branch of the upper stream of the Yangtze River. The literature recordings of this ethnic group can be traced back to the Jin Dynasty (265–420 AD; Chang, 1987, p. 210). It was still spotted in documents during the period of the Republic of China. On the other hand, "Na" is the Romanized word of the shared syllable of their endonyms. During the nationality recognition conducted by China in 1950s, the western branch of Moso People has been recognized as "Naxi" People, according to their endonym /nɑ˧ ˩ɕi˥/. According to language documentation studies, the other branches share the similar morphological structure of their endonyms: the syllable "nɑ" followed by the word for "people". The syllable "nɑ" is homophonic to the word meaning "black, big". Therefore, some scholars use "Na/Naish" People to refer to this ancient ethnic group (Jacques and Michaud, 2011). According to ISO 639–3, the dialects of Moso/Na languages are assigned to two different codes: "nxq" for Naxi and "nru" for Na.

4. The eastern branch remains still an oral community. Their writing system includes a limited number of pictographs which are only known by the Daba priests (Song, 2003; Xu, 2017a).

eign brands (e.g., "Starbucks," cf. Poupard, 2019, p. 54). Dongba priests can also apply their writings to transcribe IPA (cf. Zhao, 2013, p. 76). According to the framework of writing systems developed by Gelb (1952, pp. 190–194), Dongba script should be categorized as a semasiography used as an identifying-mnemonic device, which is undergoing the change into phonography.

The present study focuses on a crucial first step to digitize Dongba script: the construction of a font database of Dongba glyphs. The author provides an in-depth analysis of the Dongba radicals which lays a foundation for a font database. A font database includes a comprehensive collection of Dongba glyphs and sets up a semantic index that can be compared to radicals in Chinese characters.

2.   The Development of Dongba Digitization

E-Dongba is the first keyboard for typing Dongba pictograms (2001–2020). This keyboard can input Dongba pictographs through either semantics or Naxi *pinyin*. Chinese and English translations of each Dongba character are embedded into the system. The lexical database of *E-Dongba* is based on Fang and He (1981).[5] The Dongbafont implemented in *E-Dongba* includes 1,561 Dongba hieroglyphs and 661 Geba scripts, along with 50 International Phonetic Alphabet (IPA) characters for the Naxi language.

Several other attempts to type in Dongba pictographs were produced (see in Table 1). Among the five patents, three depend on semantics and two on the shapes of the characters. These keyboards aim at enabling users not knowing much about Dongba pictographs to type in these characters. The action of typing in Dongba pictographs according to their meanings involves the issue of translation. To key-in Dongba pictographs according to the composition of the characters requires a detailed analysis of several aspects of each character (components, strikes, crossing points, etc.), which could also be challenging.

Besides the virtual keyboard, several designs of Dongba script were published open-source, e.g., 遊トンパ, 遊トンパ年賀 (Kojima, 2002–2004). They are self-designed fonts in TrueType format, together with a character table displaying limited graphs (215, which is about one tenth of the inventory of Dongba pictographs).[6]

As a foundation of a key-in system, a database of the characters needs to be settled. The Unicode encoding project aims at providing an inter-

---

5. Cf. The third official Unicode proposal N4043 (L2/11–178: 2) submitted in 2011.

6. There are three ways to input text in Unicode: by selecting characters from a table; by virtual keyboard; by converting data that exist in other encodings (Haralambous, 2007, p. 159).

TABLE 1. Patents of Academic Institutions

Title and Number	Publication	Institution
纳西东巴象形文字的分类拼意输入方法及其键盘 (CN1547094A) A Keyboard to Type Dongba Pictographs Based on Semantic Catergorization	2004.11	大理学院 Dali University
东巴文图元输入法及键盘 (CN101477408A) A Keyboard to Type Dongba Pictographs Based on Structures and Strokes	2009.07	大连民族学院 Dalian Minzu Institute
纳西-汉-英输入法的实现方法 (CN103677305A) A Naxi-Chinese-English Keyboard to Type Dongba Pictographs	2014.03	昆明理工大学 Kunming University of Science and Technology
基于图形拓扑特征进行识别的纳西东巴象形文字输入方法 (CN104866117A) A Keyboard to Type Dongba Pictographs Based on Graphic Topology	2015.08	北京科技大学 University of Science and Technology Beijing
纳西东巴象形文和中文的综合输入方法 (CN106055124A) A Naxi-Chinese Keyboard to Type Dongba Pictographs	2016.10	河南农业大学 Henan Agricultural University

national standard for computer processing of the Dongba pictographs, in order to facilitate an easier transmission of the data on the internet. The preliminary task of this goal is to settle down the number of Dongba characters and the unified forms of them. The script has a tentative allocation at U+1A800 to U+1ACFF. The related proposals, comments, and meeting records are chronologically listed on the website SCRIPT-SOURCE.[7] A summary of the Unicode proposal progress is displayed in Table 2.

TABLE 2. Summary of Unicode Proposal Progress

Title (Year)	# ch.	Principles of Selecting	Order of Characters	Information of Characters
L2/00−048 (2000)	48	different syllables	alphabetic	column-row

7. The webpage of Dongba script: `https://scriptsource.org/cms/scripts/page.php?item_id=entry_detail\&uid=lb7t8h9k6v`.

N3425 (2008)	1203	A. exclude variants B. exclude most phonetic loans C. include only the stable compounds D. exclude Geba script	semantic (reference unclear, possibly *Naxi Dongba Guji Yizhu Quanji*)	serial number
N3442 (2008)	Geba signs should be included; lexical sources besides *Naxi Dongba Guji Yizhu Quanji* should be cross-checked; the characters should be assigned to tentative code points			
N3935 (2010)	1188	A, C, & D B. include only one phonetic loan	alphabetic (based on the Naxi Pinyin orthography)	English translation
N3965 (2010)	Naxi Dongba pictographs were "pre-writing" rather than a writing system.			
N4043 (2011); N4633 (2014)[8]	1188	A. exclude variants B. exclude Geba script	alphabetic (based on the Naxi Pinyin orthography)	English translation; character name (Naxi Pinyin transcription)
N4641 (2014)	provide details of each Dongba glyph, including: 1) glyphs in the Dongba script, 2) IPA transcription, 3) Romanized orthography, 4) Chinese gloss for each syllable, 5) English translation of each Chinese gloss			
N4877 (2017)	1572	Fang and He (1981)	semantic	character name (Naxi Pinyin transcription)
N4878 (2017)	2164	Li, Zhang, and He (1972)	semantic	character name (Naxi Pinyin transcription)
N4895 & N4898 (2017)	1188	the same repertoire as N4043	alphabetic	character name (Naxi Pinyin transcription); IPA transcription; gloss in both English and Chinese
N4895 (2017)	provide an explanation of how some characters can combine			

---

8. The proposal of 2014 remains the same as the 2011 one, with a supplementary file about the modern use of the Dongba script. The comments on N4043 (2011) remain the same as the ones in N3965, issued in 2010 (cf. N4060).

L2/18−321 (2018)	the repertoire of 1188 characters omitted many commonly used ligatures (onomastics); the records of polyphone and polymorphic characters; three popular methods of combination of Dongba graphemes: *combination of two ideograms attaching phonetic grapheme to a logograph adding a morphemic grapheme indicating the colour of the object*
L2/19−173 (2019)	what two-dimensional model will be used to present the script

3.   A Semantic Index for a Dongba Script Database

3.1.   The Notion of Dongba Radical

The latest Unicode Proposal arranges the Dongba glyphs following a phonetic order, the same as in Rock (1963). However, the semantic order strategy was the more popular criterion applied in Dongba dictionaries, such as Li, Zhang, and He (1972), Fang and He (1981), and Rock (1972). The 2,120 entries in Li, Zhang, and He (1972) were divided into eighteen categories. The 1,340 entries in Fang and He (1981) were divided into eighteen categories either. Rock (1972) contains fourteen categories of special vocabulary of Dongba culture, in addition to the basic vocabulary recorded in Rock (1963).

The idea substantiating Li, Zhang, and He (1972) and Fang and He (1981) is comparable to the strategy of some traditional Chinese dictionaries. For example, *Erya* 爾雅, a first surviving Chinese dictionary (pre-Qin Dynasty, cf. Karlgen, 1931, p. 49; Needham, Lu, and Huang, 1986, p. 191), which provided lexicographic guides to the Chinese characters of nineteen semantic categories. While Rock (1963) and Rock (1972) collect lexicons rather than characters.

As for a font database, *Shuowen Jiezi* 説文解字 (completed in 100 AD by Xu Shen) could be a model. In this dictionary, 9,553 Chinese characters (and other 1,163 variants) are categorized into 540 sections according to their major semantic components. The same categories are then catalogued according to the similarities of their forms.

The semantic order has been chosen partially due to the fact that the pictographic Dongba script consists of logographs. The characters with related meanings share the same graphemes, which are comparable to radicals in Chinese. A Chinese radical is a graphical component of a character "giving in a very general way something of the meaning of the character" (Chao 1948: 104−105). The notion was introduced by Xu Shen as "section" (" 部 ") in *Shuowen Jiezi*. It then became a standard in the compilation of dictionaries.[9] The term "radical" (" 部首 ") was first used

---

9. In this dictionary, Xu Shen established the six types of composition of Chinese characters (i.e., philological theory called "Liushu 六书 "). Similarly, there were dis-

in the *Kangxi Dictionary* 康熙字典 (Wilkinson, 2013, p. 74), which literally means "section heading" (Woon, 1987, pp. 147–148). The "radicals" are simplified categorizations of "sections". The number of radicals is 214 in the *Kangxi Dictionary*. The term has various other translated names, such as "semantic element," "key," "classifier," "determinative," and "signific," while "radical" is the most common (DeFrancis, 1984, p. 80).

At the current stage, none of the current key-in systems is designed according to radicals, while radicals can be highlighted among Dongba glyphs. As for the pictographic writing system, the radicals are closely related to the semantic categories of Dongba glyphs. For example, the glyphs in the category "astronomy" can be classified into fifteen groups according to their shared components: ⌒ "sky" (phonetic loan variant: ⌁ "mushroom"), ⊕ "sun," ☡ "moon," ♪ "month," ⁚⁚ "star," ✝ "star; bright," ⌁ "the constellation of dzo (hybrid of yak and cattle)," ⹀ "wind," ⪙ "rain," ⚶ "snow," ∾∾ "cloud," ⪙ "dew," ℉ "lightening," ⋀ "rainbow," and ⫯ "time".

The glyphs may contain more than one radical. For example, in the glyphs representing the four seasons, each glyph consists of the pictogram of the representative feature of the season: ⹀ "wind," ⪙ "rain," "crop on the ground" (⚘ "flower" is homophonic to the word "crop"), ⚶ "snow," along with ⁚⁚ "three" and ♪ "month".

TABLE 3. Dongba Glyphs of the Three Months of Each Season

Glyph	Naxi Pinyin	Morphemes	Gloss
	/nioq seel hei/	spring-three-month	"three months of spring"
	/roq seel hei/	summer-three-month	"three months of summer"
	/chvl seel hei/	autumn-three-month	"three months of autumn"
	/cee seel hei/	winter-three-month	"three months of winter"

Yu (2003, p. 25) categorizes Dongba glyphs into three types according to their constructions: pictogram ("单字"), ligature ("合文"), and fixed-group glyphs ("字组"). The four glyphs of the seasons mentioned above, for example, are "fixed-group glyphs". Each component corresponds to one syllable of the designation of the glyph. Among

---

cussions on the composition methods of Dongba glyphs, e.g., Z. He (1976) highlighted seven categories, Fang and He (1981, pp. 56–72) analysed ten types of structures attested in Dongba pictographs, Wang (1988, pp. 44–54) proposed five basic methods, Zhou (1994) and Yu (2008, pp. 12–37) applied the *Liushu* Theory of Chinese to Dongba script.

these components, "wind," "rain," "snow," "three," and "month," are pictograms, read as "spring," "summer," "winter," "three," and "month," respectively, while "autumn" is a ligature, which consists of two pictograms ("flower" and "ground"), but is read as the word for "autumn".

Zamblera (2018) is an initial work on a pictogram-based semantic index of Dongba writing.[10] The author categorizes the Dongba glyphs into "basic pictographs" and "complex pictographs". A "basic pictograph" is defined as "made by one and just one iconographic unit (signifier) that conveys its meaning through its pictorial resemblance to the physical object meant," while a "complex pictograph" can be divided into two sets: "composed units" and "fusion units". Table 4 shows the correspondences among the terms defined by Yu (2003) and Zamblera (2018) for the different types of Dongba glyphs.[11]

TABLE 4. The Terms for Dongba Glyph Types

Yu (2003)	Zamblera (2018)
pictogram (" 单字 ")	basic pictographs
ligature (" 合文 ")	complex pictographs—fusion units
fixed-group glyphs (" 字组 ")	complex pictographs—composed units

The notion "basic pictograph" can be a counterpart of "radical" in the Chinese writing system, while "composed units" and "fusion units" are two major methods of composition of Dongba ligatures. In Dongba writing system, a radical can be an ideogram, a phonetic symbol, or a signifier.

The function of Dongba radicals are slightly different from those in the Chinese writing, due to the grammatological features of the Dongba script. The fixed-group glyphs consist of multiple pictograms to represent a word or phrase, which are read as multiple syllables. Therefore, a pictogram in a ligature functions as the radical (in the Chinese writing), while, in a fixed-group glyph, the pictograms are components, which can be linked to the radical index.

---

10. Description of this project, references, and other documentation can be found at the webpage "Naxi Dongba: Naxi People Culture and Dongba Tradition," published by Stefano Zamblera in 2009. URL: `www.xiulong.it/4.0/Dongba/homeengl.htm`.

11. A description of these distinctive components can lead to the refinement of the two types of "complex pictographs," as in Chinese "Liushu" Theory and other discussions about the composition of Dongba glyphs.

3.2.   The Elicitation of Dongba Radicals

The present study provides an in-depth analysis of Dongba radicals based on Li, Zhang, and He (1972). The glyphs in the respective categories are ordered according to their semantic similarities. Some categories are organized more coherently to the shared component among the glyphs. Like in the case of the glyphs of the categories astronomy and geography, where each star or planet and each type of landscape, have a distinctive pictograph (cf. Appendix).

Conversely, some categories require more analysis in order to elicit the radicals. For example, the 341 glyphs (No. 230–570) in the category "humanity" generally share a basic grapheme indicating the notion of "human". However, one radical in itself would not be enough for searching purposes. Looking through the glyphs, it is possible to classify the words into nouns (natures and relations of human being) and verbs (movements of human being).

The noun section include 13 minor semantic groups: ⟨glyph⟩ "human," ethnic groups based on standing figures like ⟨glyph⟩ "Guzo People" and ⟨glyph⟩ "Bai People," ⟨glyph⟩ "human," ⟨glyph⟩ "enemy," ⟨glyph⟩-1 "woman," ⟨glyph⟩ "to sit (siting figures)," ⟨glyph⟩ "Indian," ⟨glyph⟩ "manager," ⟨glyph⟩ "dead," ⟨glyph⟩ "body," ⟨glyph⟩ "big," ⟨glyph⟩ "open hair". The glyph ⟨glyph⟩ "to sit (siting figures)" stands for a verb by itself. It has been listed in the noun section due to the fact that the glyphs connected with this radical are nouns, such as ⟨glyph⟩ "emperor," ⟨glyph⟩ "guest," and ⟨glyph⟩ "lama priest". These glyphs are created by depicting sitting figures.

The verb section include 27 minor semantic groups, describing 15 types of movements:

 1. movements of the legs: ⟨glyph⟩ "to jump," ⟨glyph⟩ "to kneel," ⟨glyph⟩ "to lean on," ⟨glyph⟩ "to fall," ⟨glyph⟩ "to tremble," ⟨glyph⟩ "to run," ⟨glyph⟩ "to step across," ⟨glyph⟩ "to fold," ⟨glyph⟩ "to get up";
 2. movements of the arms: ⟨glyph⟩ "to lift," ⟨glyph⟩ "to fly";
 3. movements of excretion: ⟨glyph⟩ "to wee";
 4. movements of the mouth: ⟨glyph⟩ "to speak";
 5. movements of creeping: ⟨glyph⟩ "to crawl";
 6. movements of the waist: ⟨glyph⟩ "to bow";
 7. movements of the head: ⟨glyph⟩ "to hit with head";
 8. movements of the hand: ⟨glyph⟩ "to paste," ⟨glyph⟩ "to bring";
 9. movements of bringing instrument: ⟨glyph⟩ "to bring knife," ⟨glyph⟩ "to pull," ⟨glyph⟩ "to chase";
10. movements of the eyes: ⟨glyph⟩ "to look";
11. movements of a people above an object: ⟨glyph⟩ "to ride";
12. movements of mind: ⟨glyph⟩ "sad";
13. movements of woman: ⟨glyph⟩-2 "woman";
14. mutual movements between two people: ⟨glyph⟩ "to fight";
15. movements of one people on another: ⟨glyph⟩ "to beat".

In Zamblera (2018), Dongba pictographs are documented according to twenty-five semantic categories with 405 subcategories ("basic pictographs").[12] The references of this dictionary include Rock (1963), Rock (1972), Fang and He (1981), Li, Zhang, and He (1972), and two recent collections: P. He (2004) and Dragan (2005).[13] While the traditional dictionaries start from "astronomy," "geography," human related aspects, and, then, all sorts of animals, etc., Zamblera (2018) reorganized them according to an order that begins from "human" and proceeds to "deities," all sorts of animals, plants, etc. Moreover, "woman" has been distinguished as an independent category, in contrast to "man," in this new work.

The number of "basic pictograph" in each semantic category can be further studied. For example, according to my analysis based on Li, Zhang, and He (1972), there are 38 radicals belonging to "body parts of human," while in the index plate of Zamblera (2018) their number is 15. By comparing Zamblera's and the author's results, thirteen radicals result to be the same. Among the two Zamblera's basic pictographs not included in my analysed radicals, "body" and "hair" are not recorded in the fourth section "parts of human body and their movements" in Li, Zhang, and He (1972). The word "body" is recorded in the third section "Humanity": No.265, 舟, /gv˦/. The word "hair" is recorded in the ninth section "Food": No.1321, ᴂ, /tsʻɛ˥/ or /kv˦tsʻɛ˥/. Cf. Li, Zhang, and He (ibid., pp. 25, 103). In contrast, "head" is a basic pictograph, incorporating "ear" and "nose" in Zamblera's plate. The other twenty-five radicals

---

12. Excerpt from Zamblera's text: "A. Man and his Occupations (4); B. Woman and her Occupations (4); C. Anthropomorphic Deities (16); D. Parts of the Human Body (15); E. Mammals (30); F. Parts of Mammals (66); G. Birds (11); H. Parts of Birds (22); I. Amphibious, Reptiles, etc. (16); K. Fish and Parts of Fish (1); L. Invertebrates and lesser Animals (12); M. Trees and Plants (20); N. Sky, Earth, Water (31); O. Buildings, Parts of Buildings (13); P. Ships and Parts of Ships (1); Q. Domestics and Funerary Furniture (4); R. Temple Furniture and Sacred Emblems (14); S. Crowns, Dress, Staves, etc. (10); T. Warfare, Hunting, Butchery (15); U. Agriculture, Crafts, and Professions (30); V. Rope, Fiber, Baskets, Bags, etc. (5); W. Vessels of Stone and Earthenware (4); X. Loaves and Cakes (6); Y. Writings, Games, Music (19); Z. Strokes, Signs derived from Tibetan, Chinese, Geba, etc., Geometrical Figures Circular, Rounded Curve, Triangle, Squared, Orthogonal (36)."

13. P. He (2004) collected 1,000 commonly used Dongba glyphs, presented in 24 semantic categories. The twenty-four categories (and their English translations) of P. He (ibid.): weather 天象, seasons 时令, geography 地理, direction and position 方位, plants 植物, birds 鸟类, beasts 兽类, insects and others 虫鱼, domestic animals 畜禽, human body 人体类, family and people 人物类, behaviour 行为类, religion 宗教, gods and ghosts 神鬼, labour 劳作, food and kitchen equipment 餐饮, housing and village 房舍, clothes 服饰, war 战争, tools and wares 器物, culture 文艺, modify 形容, state 状态, numbers and others 数词和其他词. Dragan (2005) recorded 1803 Dongba pictograms and three Dongba manuscripts translated into Serbian and one Serbian poem translated into Dongba verses.

not picked as basic pictographs in Zamblera's plate are not spotted in the character list of the Section D "Parts of Human Body" (Zamblera, 2018, pp. 56–59).[14]

Among the 38 radicals of body parts, six are "ligature radicals": ⚶ "one eye (cl.)," ⚶ "see," ⚶ "tusk," ⚶ "molar," ⚶ "to think," and ⚶ "lung," while the other 32 are simple pictograph radicals. The ligature radicals are semantically connected with their corresponding simpler pictograph radicals. For example, "one eye (cl.)" and "see" are related to ⚶ "eye," "tusk" and "molar" are related to "mouth," due to the shared form, while "to think" and "lung" are related to ⚶ "heart".[15] The additional dots in "to think" (compared to "heart") are a commonly used indicator in Dongba writing, which means "something". These ligatures are considered radicals because they are used repetitively as entities to represent a morpheme in ligatures or fixed-group glyphs.

Table 5 shows the two sets of radicals attested in the category of plants.[16] The radicals in each line are semantically connected. There are 20 basic pictographs of trees and plants in Zamblera's categorization, while they are 49 in my analysis. These elicited glyphs can be classified into 11 groups, namely: "wood," "tree," "flower," "seed," "panicle," "thorn," "leaf," "spicy (n.)," "vegetable," "grass," and "bamboo".

The ligature radicals are highlighted in blue. For example, ⚶ "liquid medicine" belongs to the radical ⚶ "flower" and it appears as a component in several ligatures, such as ⚶ "to give liquid medicine to spirits," ⚶ "to mix," and ⚶ "to separate the herb and poison". Therefore, ⚶ "liquid medicine" is picked out as a radical.

The two glyphs of "tree" and "firewood" are interchangeable, i.e.,: both ⚶ and ⚶ (⚶)can mean "tree" or "firewood". However, they should be distinct as two radicals. In the context of the form of a glyph, one pictograph is a tree standing and the other one is a tree lying. The glyphs containing these two components can be distinguished according to their pronunciations. Through a phonetic analysis, the meaning of the words, rather than the form of the components, will contribute to figure out the conditions of the "interchange" between these two pictographs.

Moreover, the glyph ⚶ can be, in fact, analyzed as two radicals: ⚶-1 "wood," a general term for wooden objects; and ⚶-2 "tree," a general term for all kinds of trees. This divergence of the glyph "tree" is an addition to Zamblera's list, aiming at a more precise semantic segmentation of the tree-related glyphs.

---

14. The Dongba font used in the article is BabelStone Naxi LLC (published in May 2017). URL: `www.babelstone.co.uk`.

15. The glyph "mouth" is missing in the BabelStone Naxi LLC font list.

16. "Z." and "X." represent the surnames of the authors, and "gl." stands for "gloss". The uncertain glosses are marked in grey.

On the other hand, among the glyphs depicting all varieties of trees, it is possible to highlight two major patterns of the glyphs: five-branch trees and three-branch trees. Two of the prototypes are chosen as radicals: (1) ⍦ "chestnut tree"[17] and (2) � "willow". The fir tree has glyphs both with three branches (⍦) and five branches (⍦). In general, the trees with stronger branches tend to be depicted with five branches (e.g., "oak" and "white birch") and the trees with thinner branches with three (e.g., "willow" and "camphor tree"). The three-branch pattern is also attested among the glyphs of thorny and herbaceous plants.

TABLE 5. Radicals of Plants

Z.	⍦	—	⍦	—		
X.	⍦-1	⍦	⍦ (⍦)	⍦		
gl.	wood	broken	firewood	one trunk of wood		
Z.	—	—	—	—	—	—
X.	⍦-2	⍦	⍦	⍦	⍦	⍦ (⍦)
gl.	tree	pine tree	cypress	chestnut tree	willow[18]	fir
Z.	⍦	—	—	—	⍦	⍦
X.	⍦	⍦	⍦	⍦	—	-flower-[19]
gl.	flower	beautiful	liquid drug	poison	turquoise; jade flower	sunflower
Z.	—	—				
X.	⍦	⍦				
gl.	seed	pod				
Z.	—	—	⍦	—	⍦	⍦
X.	⍦	⍦	⍦	⍦	⍦	⍦
gl.	panicle	rice[20]	wheat	hulled barley	millet	amaranth
Z.	—	⍦	—	⍦		
X.	⍦	⍦	⍦	—		
gl.	a kind of thorn	thorn	thorns	unknown		

---

17. This glyph is translated as "oak" in Rock (1963, p. 258).
18. "Willow" is a homophonic word of "enemy".
19. This radical is listed after the radical "flower," according to my current analysis.
20. "Rice" is a homophonic word of "to feed; human".

Z.	—	—	—	—	
X.	(glyph)	(glyph)	(glyph)	(glyph)	
gl.	leaf	tobacco leaf	soybean	hemp	

Z.	(glyph)	—	—	(glyph)	—	(glyph)
X.	(glyph)	(glyph)	(glyph)	(glyph)	(glyph)	(glyph)
gl.	cactus	zantho-xylum	ginger[21]	garlic[22]	perilla	prosperous

Z.	(glyph)	—	—	—	—	(glyph)	—
X.	(glyph), (glyph)	(glyph)	(glyph)	(glyph)	(glyph)	(glyph)	(glyph)
gl.	bracken	agar-agar	mushroom	melon	gourd	turnip (veg-etable)	chives

Z.	(glyph)	—	(glyph)	—	—	(glyph)	—
X.	(glyph)	(glyph)	(glyph)	(glyph) ((glyph), (glyph))	—	[thatch]	(glyph)
gl.	grass (gener-ation; time)	sabaigrass	mug wort	medick (god of livestock; livestock; ruminant)	farges de-caisnea fruit[23]	[thatch]	creeping wood sorrel

Z.	—	(glyph)	(glyph)
X.	(glyph)	(glyph)	—
gl.	branched horsetail [banana]	bamboo	

The differences spotted in the two sets of radicals depend partially on whether to include the ligatures in the radical system or not, and are partially due to the format of the sources quoted for the elicitation of radicals. At the current stage, the author elicits repetitively-used ligatures as radicals for the semantic index of the Dongba Script database. Moreover, it is possible to highlight some features of these radicals helping to understand the early stages of logographic writing systems.

First among them is the variety in the form of a glyph. A word may be written through several variants. The word "rainbow" can be either written through a pictogram (glyph), or (glyph), which is with an additional component "ground" ((glyph)) to indicate the location of a rainbow. It is also possible to use a phonetic loan to represent a radical. As attested in the category "Astronomy," the radical "mushroom" ((glyph) /mol/) is an occasional substitution of the radical "sky" ((glyph) /mee/), despite their different pronunciations: (glyph) "a white chaos of the initial world" (No.19), (glyph) "a black

---

21. "Ginger" is a homophonic word to "to throw".
22. "Garlic" is a homophonic word to "can" and "unit".
23. The glyph can be a phonetic loan of the word "time" (read as /dʐɯ³¹/).

chaos of the initial world" (No.20). The direction of the glyphs is flexible; however, in some cases, it may indicate different meanings. For example, "fir" can be written either as 𐂩 or 𐂪. The glyph ⌣ means "moon" when it is horizontal (⌣), and "month" when it is vertical (𐂫).

A second element is the variety in the meaning of a glyph. One glyph may represent two semes and, therefore, two radicals. For example, the pictogram ɑ̆𐂬 can be a radical meaning of either "tree" or "wood". Moreover, it uses pine tree as the prototype of trees.

As a third aspect, the radicals may lie within the glyphs, and yet they are not used analytically. For example, the glyphs 𐂭 "die" (No. 291) and 𐂮 "lie down" (No. 306) share the concept of an individual lying down. However, there is not a single character in which only a figure of an individual lying down exists without indicators. For another instance, as mentioned, the glyphs of plants are often depicted by three branches or five branches. Nevertheless, there is not a glyph representing the general notion of three-branch plant or five-branch plant.

4.   Latin Alphabets for Semantic Index

Multiple indexes, including phonetic index and radical index, are available in contemporary dictionaries (e.g., *Xinhua Dictionary*). Li, Zhang, and He (1972) provided indexes according to the IPA transcription and stroke numbers of Chinese glosses of each Dongba glyph. In other words, the alphabetic order and semantic order do not exclude each other.

For the character names, it is necessary to set up an orthography for transcribing IPA into Naxi *pinyin*. Naxi *pinyin* is designed for the standard Naxi language (Dayan dialect). Its Romanized letters can unify the various sets of symbols applied to transcribe Dongba glyphs pronunciations.[24]

However, Naxi *pinyin* needs to expand the inventory of notations in order to transcribe other dialects of Moso People. According to He and Jiang (1985, pp. 130−133), there is no contrast between voiced consonants and nasalized consonants, velar and uvular consonants, dental and retroflex stops. Therefore, nasalized consonantal initials, such as /mb/, /nd/, /ŋg/, are transcribed as /bb/, /dd/, /gg/, and no symbols for voiced consonantal initials exist. There are no symbols for neither uvular nor

---

24. Naxi *pinyin* orthography: /p/, /pʰ/, /b/, /m/, /f/, /t/, /tʰ/, /d/, /n/, /l/, /k/, /kʰ/, /g/, /ŋ/, /h/, /tɕ/, /tɕʰ/, /dʑ/, /ɕ/, /tʂ/, /tʂʰ/, /dʐ/, /ʂ/, /ʐ/, /ts/, /tsʰ/, /dz/, /s/, /z/, /i/, /y/, /æ/, /ɑ/, /o/, /u/, /v/, /ə/, /ər/, /iə/, /uæ/, /uɑ/, /uə/. As for a concise key-in convention, in N4877 and N4878, the letter /y/ stands for /ʏ/ in Li, Zhang, and He (1972), /a/ for /ɑ/, /e/ for /ɛ/, and /ə/ for /ʌ/.

retroflex consonantal initials since these phonetic values appear in conditional contexts. Nevertheless, minimal pairs of these consonantal initials exist in dialects of the Naish languages. For example, the contrasts between voiced initials and nasalized initials are attested in Li, Zhang, and He (1972), whose phonemic system reflects one of the western dialects in Ludian area. For another instance, Ruke People, an intermediate ethnic group of Moso People, distinguish voiced consonants and nasalized consonants and dental and retroflex stops. Ruke Dongba culture preserves rituals different from the Naxi Dongba context (He and Guo, 1985, p. 40). Dongba glyphs of Ruke People are recorded as an independent category in Li, Zhang, and He (1972, pp. 125–127). It is possible that the repertoire of Dongba pictographs will incorporate the Ruke Dongba pictographs transcribed in their phonemic system.

A more complete Latin alphabet for Naxi can be found on the "Omniglot" website, including symbols for both voiced consonants and nasalized consonants, edited by Michael Peter Füstumum and Wolfram Siegel.[25] Another main addition to the Omniglot version would be the retroflex initials. The author suggests inserting an "r" after the alveolar initials to stand for their retroflex counterparts. Letter "r" already exists in the current Latin alphabet for Naxi. It is used as an initial and, sometimes, trills, when it appears in the end of the syllable. Its usage to indicate a retroflex is not confused with its other functions and does not increase the number of letters in Naxi *pinyin*. Table 6 displays an expanded inventory of Latin alphabets for transcribing Naish languages. In comparison to the current Naxi *pinyin*, the modifications by Omniglot are highlighted in orange and the additions by the author are highlighted in green.

TABLE 6. Expanded Latin Alphabet for Naish Languages

Initials								
	stop				nasal	approx.	fricative	
labial	p[pʻ]	b[p]	bb[b]	nb[mb]	m[m]		f[f]	v[v]
dental	t[tʻ]	d[t]	dd[d]	nd[nd]	n[n]	l[l] (/r/:[ɾ])[26]		
velar	k[kʻ]	g[k]	gg[g]	mg[ŋg]	ng[ŋ]		h[x]([h])	w[ɣ]([ʁ])
retroflex	tr[ʈʻ]	dr[ʈ]	ddr[ɖ]	ndr[ɳɖ]	nr[ɳ]	lr[ɭ]		
	affricate							
dental	c[tsʻ]	z[ts]	zz[dz]	nz[ndz]			s[s]	ss[z][27]
alveo-palatal	q[tɕʻ]	j[tɕ]	jj[dʑ]	nj[ndʑ]	ni[n̠]		x[ɕ]	y[z]
retroflex	ch[tʂʻ]	zh[tʂ]	rh[dʐ]	nr[ndʐ]			sh[ʂ]	r[ʐ]

Glides								
u[w]	i[j]							

Rhymes								
i [i]([ɿ],[ʅ])	u[u]	iu[y]	ei[e]	ai[æ]	a[ɑ]	o[o]	e[ə]	er[ɚ]
ee[ɯ]	v[ɣ]							

Tones								
-l: high	-: mid	-q: low	-f: rising					

5. Conclusion

The present study reviewed the achievements of the related project, including the implementations of key-in systems, fonts, and Unicode. The author has shown that a critical issue for the Unicode recognition of Dongba pictographs is to build up a comprehensive character list of this writing system. The dictionaries provide valuable sources on the Dongba script, from which a complete repertoire of Dongba glyphs can be elicited.

Most of the dictionaries apply a semantic categorization of Dongba pictographs, while a few use an alphabetical approach according to the phonetic transcriptions. The current semantic categorizations adopted by Dongba Dictionaries can be refined to a more detailed semantic index. Glyphs sharing one common pictograph and semantically connected with each other can be grouped as one subcategory.

The pictograph, the semantic component giving the core idea of the glyph, may correspond to the notion of "radical". Similarly to the early Chinese dictionary *Shuowen Jiezi*, a Dongba radical can be either a simple pictograph or a ligature. Each radical represents an entry for the semantic index. Such semantic index is conventionally implemented into the Chinese dictionaries.

The elicitation of radicals reveals the features of similar markers in pictographic writing systems. Through an in-depth examination of the shared components among Dongba glyphs, the author pointed out three types of 'loose' correspondences between their forms and meanings. Futher on, this research can also contribute to our understanding of how similar markers arose in mature logographic systems like

---

26. The phoneme /r/ is elicited from Li, Zhang, and He (1972). The phonetic value is [ɾ] as an initial. If this symbol is at the end of a rhyme, it indicates that the rhyme is a retroflexive vowel (ibid., p. XXIII).

27. Some dialects have nasalized voiced fricative initials. In those cases, it can be transcribed as /nss/.

Chinese logographs, Egyptian hieroglyphs, and Mayan glyphs (Handel, 2017, p. 3).

Other indexes, such as the alphabetical index of the phonetic transcriptions or English glosses and the number of strokes of their Chinese glosses, are not excluded by the semantic index. Moreover, multiple indexes could allow searching for glyphs through different channels. The major radical and related radical provide more paths to get to one glyph and, therefore, more analysis' options to the grammatological feature of one glyph. The alphabetical index depends on the Naxi *pinyin* transcription, which requires more Latin letters for a possible extended character list involving dialectal vocabularies.

A. Dongba Radicals of Astronomy, Geography, Humanity, and Human Body Parts Based on Li, Zhang, and He (1972)

(The English terms are quoted from Li, Zhang, and He, 2001, pp. 419–421, translated by Tseng Chao-yueh.)

A.1. Phenomena Connected With Heaven 天文類 (15/117)

/mee/	/nii/	/hei mei/	/hei/	/geeq/	/mbo/	/ssiuq/
sky	sun	moon	month	star	star; bright	constellation of dzo (hybrid of yak and cattle)
/hai/	/heeq/	/mbei/	/jiq/	/nrur/	/ciul/	/mi xil jjiq teeq/
wind	rain	snow	cloud	dew	lightening	rainbow
/rheeq/						
time						

A.2. Geographical Phases 地理類 (13/112)

/ddiuq/	/njoq/	/jjoqnar-uallua/	/mboq/	lv	aiq
ground	mountain	scared moun-tain	slope	stone	cliff

/jjiq/	/tel/	/chv/	/heel/	/ree/	/lee ddraiq/	rurq
water	drop	nitrate	lake	road	field	corner

A.3. Human Natures, Relations, and Movements 人文類 (40/341)

A.3.1. 人文類-1: Natures and Relations (13)

/xi/	/ggv zzeeq/	/lei bbv/	/coq/	/rvq/	/mil/
human	Guzo People	Bai People	human	enemy	woman

/nzeeq/ to sit (sitting figures)	/jje aq/ Indian	/nzee/ manager	/shee/ dead	ggv body	/ddeeq/ big	/baq/ open hair

A.3.2. 人文類-2: Movements (27)

/co/ to jump	/ceel/ to kneel	/toq/ to lean on	/ndol/ to fall	/niol/ to tremble	/xa/ to run	/gua/ to step across
/sseel/ to fold	/dee/ to get up	/lvq/ to lift	/nziq/ to fly	/mbi/ to wee	/lerq/ to speak	/mbvq/ to climb
/ggvq/ to bow	/tail/ to hit with head	/bal/ to paste	/bol/ to bring	/hai/ to bring knife	/daiq/ to pull	/ndiul/ to chase

/liuq/	/nrai/	/nvl mei qil/	/mil/	/aiq/	/lal/
to look	to ride	sad	woman	to fight	to beat

A.4. Parts of the Human Body and Their Movements 人體類 (38/115)

/kv/	/mieq/,/nieq//pol/	/ddoq/	/hei/	/niil merq/	/nvl/
head	eye	one eye (cl.)	see	ear	nose	mouth

/beiq/	/nraiq/	/mgoq/	/hee/	/sal/	/lua baq/	/gei/
to spit	tusk	molar	teeth	breath	beard	neck

/laq/	/niil nii/	/piq/	/hoq/	/nee/,/nvl mei/	/shv lv/	/churl/
hand	breast	shoulder blade	rib	heart	to think	lung

/guaq/	/hol/	/zil/	/bbv/	/geeq/	/sel/	/bbiu liu/
diaphragm	stomach	pancreas	intestine	gut	liver	kidney

/teel/ lumbar spine	/shai/ blood	/mgv/ tendon	/shee/ meat	/cel/ misfortune	/ua/ bone	/lai/ penis

/pv/	/mei/	/kee/
male	female	foot

References

Chang, Qu [常璩] (1987). 华阳国志校补图注 *[A Correction and Completion with Pictographic Annotation for Local Chronicles of South China]*. Ed. by Ren Naiqiang [任乃强]. 上海 [Shanghai]: 上海古籍出版社 [Shanghai Guji Chubanshe].

Chao, Yuen-Ren (1968). *Language and Symbolic Systems*. Cambridge, UK: Cambridge University Press.

DeFrancis, John (1984). *The Chinese Language: Fact and Fantasy*. Honolulu: University of Hawai'i Press.

Dongba Culture Research Institute [东巴文化研究院] (1999). 纳西东巴古籍译注全集 *[An Annotated Collection of Naxi Dongba Manuscripts]*. 昆明 [Kunming]: 云南人民出版社 [Yunnan Renmin Chubanshe].

Dragan, Janekovic（2005）. *Na-si: srpski recnik [Nakhi-Serbian Dictionary]*. Beograd: Narodna biblioteka Srbije [National Library of Serbia].

Fang, Guoyu [方国瑜] and Zhiwu He [和志武] (1981). 纳西象形文字谱 *[A Dictionary of the Naxi Pictographic Writing]*. 昆明 [Kunming]: 云南人民出版社 [Yunnan Renmin Chubanshe].

Fu, Maoji [傅懋勣] (1982). "纳西族图画文字和象形文字的区别 [The Difference between Pictograms and Hieroglyphs of the Naxi People]." In: 民族语文 *[National Language]* 1, pp. 1–9.

Gelb, Ignace（1952）. *A Study of Writing*. Chicago: The University of Chicago Press.

Guo, Pu [郭璞] (1999). 爾雅注疏 *[Song Dynasty, Commentary and Subcommentary of Erya]*. 北京 [Beijing]: 北京大学出版社 [Beijing University Press].

Handel, Zev (2017). "The Cognitive Role of Semantic Classifiers in Modern Chinese Writing as Reflected in Neogram Creation." In: *Seen, not Heard: Composition, Iconicity, and the Classifier Systems of Logosyllabic Scripts*. Ed. by Ilona Zsolnay. to appear.

Haralambous, Yannis (2007). *Fonts & Encodings*. Sebastopol: O'Reilly.

He, Jiren [和即仁] and Zhuyi Jiang [姜竹仪] (1985). 纳西语简志 *[A Brief Description of the Naxi Language]*. 北京 [Beijing]: 民族出版社 [Minzu Chubanshe].

He, Pinzheng [和品正] (2004). 东巴常用字典 *[Naxi Dongba Pictographic Dictionary]*. 昆明 [Kunming]: 云南美术出版社 [Yunnan Meishu Chubanshe].

He, Zhiwu [和志武] (1976). 纳西族的象形文字和东巴经 *[The Pictographic Writing and Dongba Scripture of the Naxi People]*. 昆明 [Kunming]: 云南大学历史研究所 [Yunnan Daxue Lishi Yanjiusuo].

He, Zhiwu [和志武] and Dalie Guo [郭大烈] (1985). "东巴教的派系和现状 [Branches of Dongbaism and Their Current Status]." In: 东巴文化论集 *[Collection of Papers on Dongba Culture]*. Ed. by Dalie Guo [郭大烈] and Shiguang Yang 杨世光. 昆明 [Kunming]: 云南人民出版社 [Yunnan Renmin Chubanshe], pp. 38–44.

Howard, James (1960). "Dakota Winter Counts as a Source of Plains History." In: *Bureau of American Ethnology Bulletin* 173.61, pp. 335–416.

Institute of Linguistics, Chinese Academy of Social Sciences [中科院语言研究所] (2010). In: 新华字典 *[New Chinese Character Dictionary]*. 北京 [Beijing]: 商务印书馆 [Commercial Press].

Jacques, Guillaume and Alexis Michaud (2011). "Approaching the historical phonology of three highly eroded Sino-Tibetan languages: Naxi, Na and Laze." In: *Diachronica* 28, pp. 468–498.

Karlgen, Bernhard (1931). "The Early History of the Chou Li and Tso Chuan Texts." In: *Bulletin of the Museum of Far Eastern Antiquities* 3, pp. 1–59.

Kojima, Tomio [小島富美男] (2002–2004). "遊トンパ年賀 [Yu Tompa New Year]." URL: http://www.efword.com/tompa/font/.

Li, Lifen [李例芬] (1997). "纳西东巴古籍与语言研究 [A Study on Naxi Dongba Scripture and Language]." In: *Journal of Yunnan Institute for Nationalities* 4, pp. 69–73.

Li, Lincan [李霖灿], Kun Zhang [张琨], and Cai He [和才] (1972). 么些象形文字字典 *[A Dictionary of Mo-So Hieroglyphics]*. 台北 [Taipei]: 文史哲出版社 [Wenshizhe Publishing House].

———— (2001). 纳西族象形标音文字字典 *[Naxi Pictographic Symbols Dictionary]*. 昆明 [Kunming]: 云南民族出版社 [Yunnan Minzu Chubanshe].

Liu, Chulong [刘楚龙] and Tengyu Huang [黄滕宇] (2013). "藏族支系木雅人历书解读 [A Preliminary Interpretation and Research on Muya Tibetan's Almanacs]." In: 语言学研究 *[Linguistic Research]* 1, pp. 12–24.

Mueggler, Erik (2011). *The Paper Road: Archive and Experience in the Botanical Exploration of West China and Tibet*. Berkeley, CA: University of California Press.

Needham, Joseph, Gwei-djen Lu, and Hsing-Tsung Huang (1986). *Biology and Biological Technology. Part 1: Botany*. Vol. 6. Science and Civilisation in China. New York: Cambridge University Press.

Poupard, Duncan (2019). "Revitalising Naxi dongba as a 'pictographic' vernacular script." In: *Journal of Chinese Writing Systems* 3.1, pp. 53–67.

Ramsey, Rovert (1987). *The Languages of China*. Princeton: Princeton University Press.

Rock, Joseph (1963). *A ^{1}Na-^{2}Khi-English Encyclopedic Dictionary (Part I)*. Rome: Istituto Italiano per il Medio ed Estremo Oriente.

———— (1972). *A ^{1}Na-^{2}Khi-English Encyclopedic Dictionary (Part II)*. Rome: Istituto Italiano per il Medio ed Estremo Oriente.

Song, Zhaolin [宋兆麟] (2003). "摩梭人的象形文字 [Hieroglyphic Writing of Mosuo People]." In: *Southeast Culture* 4, pp. 86–93.

———— (2010). "西南民族象形文字链探析 [A Study on the Pictographic Writing Chain of the Ethnic Groups in South West China]." In: 民族艺术 *[National Art]* 3, pp. 31–35.

Sun, Hongkai [孙宏开] (1982). "尔苏沙巴图画文字 [Pictographic Writing of Ersu Shaba]." In: 民族语文 *[National Language]* 6, pp. 44–49.

Terrien de Lacouperie, Albert (1894). *Beginnings of Writing in Central and Eastern Asia, or Notes on 450 Embryo-Writings and Scripts*. London: Nutt.

Wang, Dehe [王德和] and Xuan Wang [王轩] (2013). "尔苏沙巴文历书中虎推地球图 [The Figure of the Tiger Pushing the Earth in the Ersu People's Hemerology Written in Shaba Script]." In: 中国藏学, *[Chinese Tibetology]* 4, pp. 148–155.

Wang, Yuanlu [王元鹿] (1988). 汉古文字与纳西东巴文字比较研究 *[A Comparative Study of Ancient Chinese Characters and Naxi Dongba Glyphs]*. 上海 [Shanghai]: 华东师范大学出版社 [Huadong Shifan Daxue Chubanshe].

Wilkinson, Endymion (2013). *Chinese History: A New Manual*. Cambridge (MA: Harvard University Asia Center.

Woon, Wee Lee [云惟利] (1987). 漢字的原始和演變 *[Chinese Writing: Its Origin and Evolution]*. Macau: University of East Asia.

Xu, Duoduo (2017a). "From Daba Script to Dongba Script: A Diachronic Exploration of the History of Moso Pictographic Writings." In: *Libellarium: Journal for the Research of Writing, Books, and Cultural Heritage Institutions* 10.1, pp. 1–47.

——— (2017b). "Phonemic and Tonal Analysis of Youmi Ruke." In: *Annals of the University of Craiova - Series Philology / Linguistics* 39, pp. 239–251.

Yu, Suisheng [喻遂生] (2003). 纳西东巴文研究丛稿 *[Miscellaneous Researches on Naxi Dongba Culture]*. 成都 [Chengdu]: 巴蜀出版社 [Bashu Shushe].

——— (2008). 纳西东巴文研究丛稿二 *[Miscellaneous Researches on Naxi Dongba Culture (II)]*. 成都 [Chengdu]: 巴蜀出版社 [Bashu Shushe].

——— (2009). "纳西东巴文献学纲要 [Outline of the Bibliography of Naxi Dongba Literature]." In: 历史文献研究 *[Historical Document Research]* 28, pp. 12–21.

Zamblera, Stefano (2009). "Naxi Dongba: Naxi People Culture and Dongba Tradition." URL: http://www.xiulong.it/4.0/Dongba/.

——— (2018). *A Dongba Pictographs Dictionary with Iconographic Index Plates*. Morrisville, NC: Lulu.com.

Zhao, Liming [赵丽明] (2013). "从宗教走向世俗、从原始走向成熟　从白地、油米、宝山东巴文书等看东巴文的两大突破 [From Religious to Secular, From Primitive to Mature—Two Breakthroughs of Dongba Script Evidenced in Baidi, Youmi, and Baoshan Data]." In: *Linguistic Research* 1, pp. 68–86.

——— (2014). "珞巴族象形符号 [Pictographic Symbols of Lhoba People]." In: 中国社会语言学 *[Chinese Sociolinguistics]* 23.2, pp. 66–72.

Zhao, Liming [赵丽明] and Zhaolin Song [宋兆麟] (2011). 中国西南濒危文字图录 *[Catalog of the Endangered Scripts in Southwest China]*. 北京 [Beijing]: 学苑出版社 [Academy Press].

Zhou, Youguang [周有光] (1994). "纳西文字中的《六书》 [The "Liushu" Theory in the Writing System of Naxi People]." In: 民族语文 *[Minority Languages of China]* 6, pp. 12–19.

Transcribing Sign Languages With TYPANNOT: The Typographic System That Retains and Displays Layers of Information

Claire Danet, Dominique Boutet†, Patrick Doan,
Claudia Savina Bianchini, Adrien Contesse, Léa Chèvrefils,
Morgane Rébulard, Chloé Thomas & Jean-François Dauphin

Abstract. There are more than 140 sign languages (SLs) in the world and studying them is a relatively recent field of research (starting in the 1960s). Linguists have the need to represent the different levels of gestures that make up the signs in order to analyze the way SLs work. Such transcription requires the use of a dedicated graphic system (Slobin et al., 2001).

TYPANNOT, the transcription system presented in this article, is a typographic system that allows the description of all formal features of SLs. Our contribution to the field of grapholinguistics is a phonological model and a transcription system for SLs that are rooted in the articulatory possibilities of the signer's body. Compared to existing graphematic systems, our approach of SLs description is both phonological, allowing descriptions of the different articulatory structures (low level) involved in SLs, and logographical, allowing users to read the transcriptions from a unified perspective (high level).

We will detail the design principles that drive the development of such a typographic system, the graphemic model that derives from linguistic study, and the tools that allow researchers to use TYPANNOT to its fullest capacities.

This article also outlines the kinesiological approach (Boutet, 2018), which TYPANNOT uses, noting radical changes in the way researchers should look at meaning through gesture. This approach opens new perspectives in researching movement itself as a central source of meaning in human communication via gesture.

Introduction

Sign languages (SL), of which there exist at least 144 worldwide (Ethnologue.com, 2020), are gestural languages with grammatical/linguistic structures based on body expression (Sandler and Lillo-Martin, 2006).

Claire Danet   ⓘ 0000-0001-5362-8924
Université de la Sorbonne & Université Technique de Compiègne
claire.danet@gmail.com
Patrick Doan ⓘ 0000-0002-2012-1630, Claudia S. Bianchini ⓘ 0000-0002-4783-1202
Adrien Contesse ⓘ 0000-0002-8997-3321, Léa Chèvrefils ⓘ 0000-0001-9123-1223
Morgane Rébulard ⓘ 0000-0002-8469-9693, Chloé Thomas ⓘ 0000-0002-8562-1249

This visuogestural modality means that SLs work in very different ways compared to vocal languages (VL). First, SLs are 4-dimensional (3 space dimensions plus time) by nature and employ spatialization, meaning that grammatical elements manifest in space like on a scene, for example verbs appear to move from the sender to the receiver. Second, SLs can articulate several gestures in parallel, using the entire body to express multiple information simultaneously (Braffort, 1996). If VLs can be seen as monolinear, meaning only one piece of information is being communicated at a time, SLs can be seen as plurilinear (Cuxac, 2001). Third, semiotically speaking, SL signs are iconic: their appearance tends to resemble some aspect of the thing or action being denoted.

Since the 1960s, with the development of linguistic research and the recognition within hearing communities that SLs are full-fledged languages, most linguists have agreed on the phonological[1] decomposition of SLs into two parameters: (a) manual, such as the configuration of the hand, its orientation in space, its location and its movement; and (b) non-manual, such as gaze, facial expression, and torso posture.

Unfortunately, this consensus didn't translate into the conception and the adoption of a unified transcription system for SLs. Until today, the multi spatial and multi parametric properties of SLs cannot be properly represented by neither a dedicated symbolic system nor a VL description system.

Nevertheless, various attempts for writing SLs do exist, e.g., in France, many historical instances can be cited, such as "Mimography" (Bébian, 1825) and "D'Sign" (Jouison, 1995), or more recently "Signography" (Haouam-Bourgeois, 2007), "Schematization" (Guitteny, 2007) and "SL-video" (Brugeille, 2007). These different graphic forms were created to compensate for the lack of traceability in situations such as teaching SL or sharing artistic expressions (poetry, sign-singing, theater pieces, etc.). Bianchini (2012) even considers that SL writing would be an additional route for hearing people to enter the Deaf world.

With the birth of SL linguistics, and in particular after the research of William Stokoe (1960), different notation systems have emerged. Some were created with the aim of detailed analysis and transcription (among them, Prillwitz et al., 1989; Stokoe, 1960; these are phonographic systems, in which each grapheme transcribes a phoneme. Other, more logographic systems (a grapheme representing an entire lemma, i.e., word), like SignWriting (Sutton, 1995) or Si5s (Augustus, Ritchie, and Stecker, 2013), represent a more functional and accessible approach for the Deaf community.

---

1. Here and throughout the article, phonological is of course referring to the phonetics and phonemics of SLs, in which visual form is abstracted into units of meaning, or phonemes, and we are not using phonological to mean the science of speech sounds.

Whatever the motivation for finding ways to represent SLs is, their visuogestural natures, their spatial and temporal dimensions, as well as their plurilinearity transform the task into a particularly daunting challenge (Boutet, 2018).

Keeping in mind those specificities, the GestualScript team at ÉSAD-Amiens[2] reviewed the existing linguistic models and graphematic systems in order to understand their underlying strengths and limitations (§ 1). This work fueled the design thinking behind the conception of TYPANNOT (§ 2), a typographic system that takes advantage of new technologies to tackle SLs representation problems while adopting a radical perspective in order to completely revisit the existing descriptive models (§ 3).

1. Existing SL Representation for Various Scopes

Although pursuing different objectives, SignWriting (SW) and HamNoSys are two examples of notation systems that both rely on a parametric approach to organize the representation of SLs. While SW aims to offer writing within the framework of teaching with SL, HamNoSys is focused on transcribing SLs in order to analyze them systematically. These two distinct perspectives yield different yet complementary principles of graphic representation.

1.1. SignWriting (SW): a Representation System for Recognizing SL

In 1974, Valerie Sutton conceived SW, inspired by her previous DanceWriting (1966-74) work and driven by the linguistic research carried out at the University of Copenhagen (Sutton, 2020). This system is aimed at both the teaching and the everyday practice of SL, and is characterized by an anthropomorphic representation of the sign in an attempt to offer a proxy of reality.

First, SL signs are represented by distinct graphic units (graphemes) that correspond to the minimum units that carry meaning in the structure of the language (phonemes), and which take up the main formal characteristics (shapes of the hand, eyes, arms, etc.); this is a so-called phonological level of deconstruction. Next, these graphemes are arranged analogically to the sign space in a thumbnail called a "vignette"

---

2. The GestualScript team, based at the De-sign-e lab of the École Supérieure d'Art et de Design (ÉSAD) d'Amiens, is an interdisciplinary group made of linguists (D. Boutet†, C. Danet, C. S. Bianchini, L. Chevrefils, C. Thomas), designers (P. Doan, M. Rébulard, A. Contesse), and a computer scientist (J.-F. Dauphin). The team's research was partly funded by the French DGLFLF and the Department of Culture.

(Fig. 1). They take up the global spatial organization of the SL sign to reproduce its image, like a transfer from reality. This is a so-called logographic level of construction where the different graphemes are brought together to form a unified sign representing a lexical unit.

This makes SW a system with a pictographic tendency since the vignettes reproduce a schematic, stylized, and above all, unified version of the SL sign, thus allowing the user to focus more on the text meaning than on the language structure. This representation, however, can also be considered "alphabetical" since each vignette can be split into glyphs which relate more to phonemes.[3]
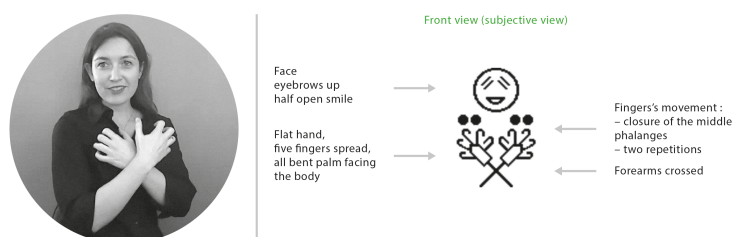
FIGURE 1. Organization of SW glyphs inside a vignette

These glyphs are then arranged non-linearly, leaving the writer considerable expressive freedom, both in the choice of several glyphs that are almost synonymous and of their location. If the only limit to this freedom is keeping the legibility of the thumbnail, this results in a great variability from one writer to another, which limits data comparisons (e.g., when searching for inter-annotator agreement) (Fig. 2).

FIGURE 2. Freedom in placing some SW glyphs (e.g., movement arrows) may make difficult data comparisons

---

3. The question of the exact nature of SW (alphabetic or featural; phonological or phonetic) is still open but not the subject of this article; for further discussion see Bianchini (2012; 2016).

SW is organized in categories and sub-categories (e.g., configuration of the hand, its movement, its dynamics and coordination, etc.); each containing basic glyphs (about 500 in total) to which specific rules apply, a process that generates nearly 37,000 "conjugated" glyphs.

The description of the elements present in a category or subcategory calls on several frames of reference (FoR, §3.1.2). For example, the movement of the hands is described in an environmental FoR (the movements are directed towards the imaginary walls of a room, the horizontal axis corresponds to the floor, the vertical one to the height of the walls; Fig. 3a), but those of the head and the body are described in a FoR centered on the speaker (the movements are directed towards the sides of the signer, the horizontal axis corresponds to the shoulders, and the vertical axis to the signer's height; Fig. 3b).

(a)(b)(c)(d)

(a) (b) (c) (d)
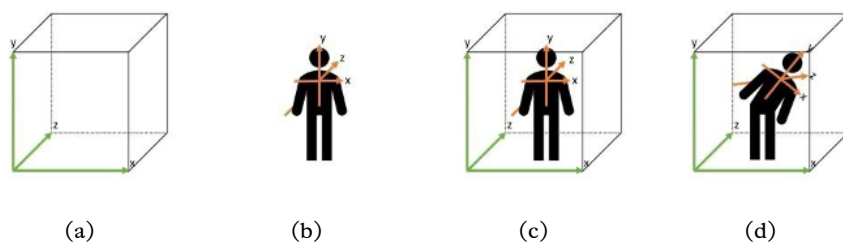
FIGURE 3. (a) the environmental FoR is used to code hand movement; (b) speaker-centered FoR is used to code head movement; if hand and head movement is coded in the same vignette, two FoRs are present. These can be: (c) collinear, if their axes are superimposed; or (d) non-collinear, involving fragmentation of the representation space

The presence of several FoRs within the same vignette generates a fragmentation of the sign representation space whenever the axes of the different FoRs are non-collinear (not superimposable). If standing, the signer's horizontal axis corresponds to the environmental horizontal axis (the two FoRs are therefore collinear; Fig. 3c), but if bent to the side, the signer's shoulders will no longer be parallel to the floor and therefore the horizontal axis will no longer correspond to the horizontal plane of the room (Fig. 3d).

The flexibility of notation and the large number of glyphs make SW an asset for representing many phenomena. Thanks to its visual evocational power, it is the system most used by educators around the world; however, in the absence of a ductus—a defined procedure specifying the number of strokes, the direction, and the sequence for draw-

ing the various symbols—writing SW is much more complex than reading SW. Moreover, this makes it difficult for linguists to obtain interannotator agreement as well as the ability to query the vignettes in a database. However, apart from the drawing of SW by hand, there is also online input software (SignMaker[4]) available and the system has been coded under the Unicode standard since 2010.

1.2.   HamNoSys: A Representation System for Analyzing SL

Directly focused on researchers, the Hamburg Notation System (Prillwitz et al., 1989), a.k.a. HamNoSys, is a transcription system based on phonological principles, i.e., each parameter is broken down into phonemes which are represented by glyphs. This approach is an evolution of the one adopted by Stokoe Notation (Stokoe, 1960), the first SL notation system.

Compared to Stokoe Notation, HamNoSys offers a more detailed description of SL phonemes and increases the number of examined parameters (e.g., non-manual parameters, locations outside the signature zone, etc.). Phonemes are represented by around 210 more or less iconic symbols, but while some have their own symbol, others are obtained by composing a basic form with diacritics (e.g., the hand configuration, the movement, etc.). This phonographic system is intended to be inter operational, and therefore aims at international use, compatibility with standard computer display and indexing tools, extension capacity, ergonomic syntax according to the principles of compositionality, syntactic efficiency (e.g., the principles of symmetry), and an iconicity of symbols (for ease of memorization) (Hanke, 2004).

The graphemic equation puts the sign information in linear order from left to right, according to a strict syntax (Fig. 4).

Like SW, HamNoSys changes its point of view. For example, the glyphs representing the hand orientation and movement are related to three perspectives, one from the signer's point of view, the others from above or from the right (Fig. 5, from ibid.).

To use HamNoSys, it is possible to download a font and a dedicated virtual keyboard. HamNoSys is also coded under the Unicode standard. SL signs are encoded in a fully linearized typographic form, which gives the system great flexibility and compatibility with computer tools for displaying and indexing data. However, HamNoSys can be complex to use, especially during the decryption phase, due to the amount of parameters to be processed and the way the characters are composed.

---

4. SignMaker (`http://www.signbank.org/signmaker.html`); for an analysis of the interface see Bianchini, 2012.
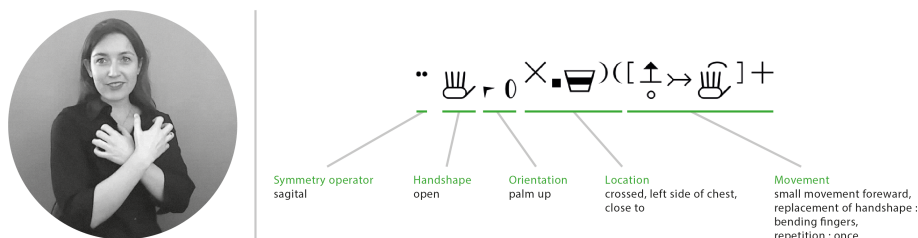
FIGURE 4. Organization of HamNoSys glyphs in an equation
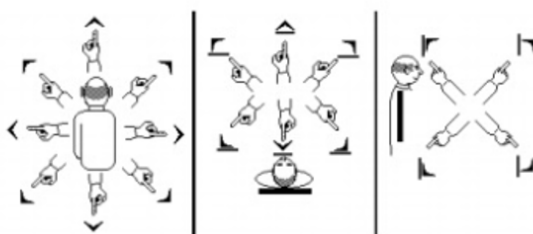
FIGURE 5. Multiple points of view within HamNoSys

1.3.   Conclusions: An External Perspective of Sl Representation

This short presentation of SW and HamNoSys has shown the advantages associated with the two main modes of representation that characterize them. The phonographic approach uses a limited inventory of signs corresponding to the SL phonological structure to transcribe them in an efficient and detailed manner; conversely, the logographic approach offers a synthetic and evocative graphic representation by visually transposing the semiotic dimensions of a corporal FoR inherent to all SLs.

It becomes clear that a semiography (linguistic sign notation that refers to the semantic level of a language) that could combine phonographic and iconic trends would be advantageous. On one hand, the phonological structuring makes it possible to isolate the distinctive elements, thus providing an efficient and functional system necessary for transcription. On the other hand, the synthetic and evocative graphic representation preserves the semiotic relationships intrinsically offered by the different spatial and bodily references of these visuogestural languages.

It is worth noting that whatever the pictographic or phonological dominant, the systems of SL representation resort in general to a form of visual iconicity. Indeed, unlike VLs which must use graphic conventions

to represent sounds, SLs can find in writing the figurative dimensions directly emanating from the visuogestural modality. Translated graphically, the articulated structure of the body and the forms it produces in space constitute an image that recalls the SL sign all at once. The result is a remarkably natural spelling, so to speak, but it remains important to ask questions about its strengths and weaknesses.

Indeed, this immediate readability means that the existing systems systematically approach the representation of SL as a projection of perceived shapes on a 2-dimensional plane. Although these languages are deeply rooted and subject to the articulators of the body and their own geometry, the body is viewed from an external point of view where only its apparent parameters are described. The problem here is not so much to have resorted to an external perspective or projection, but to have ignored the intrinsic characteristics of the SL sign. Without them, it is impossible to know what a representation really corresponds to, given the superficiality of the projection. In fact, signs are described according to what the recipient sees and not what the speaker's body is doing.

Therefore, SW and HamNoSys must multiply the points of view to account for a gestural phenomenon. These external views changing from one sign to another (HamNoSys) or even joined in the same description (SW) may lead to misunderstandings in sign reading and inconsistencies in analysis.

The linguistic distinction in manual and non-manual parameters produces dissociation between the different segments of the body, which disconnects them from their bodily transformations. Movement, which is the most reluctant parameter to be used in linguistic description due to its complexity (Boutet, 2018) is then simply considered a trajectory of the hand, yet this manual trajectory—as a trace left behind—cannot by itself contain all the meaning conveyed by the signer's bodily comportment.

2.   TYPANNOT

2.1.   Approach and Goals

The GestualScript team believes that meaning in SL is driven by the signer's own activity and that this activity is fundamentally defined by the many ways in which the body can be mobilized and experienced to promote an ongoing dynamic of signification (Poizat, Salini, and Durand, 2013; Theureau, 2004; Varela, Thompson, and Rosch, 1993). Although SL gesture refers to cultural and linguistic forms, part of its meaning is fundamentally undetermined and arises through the non-linear, open dynamics of activity. This means that gesture is personally lived and understood at the level of a body that can freely transform,

modulate, and interact with those cultural and linguistic forms within the limits of what is possible in terms of movement and signification. TYPANNOT is a novel typographic system, which allows researchers to represent those transformations, modulations, and interactions at the articulatory level, at the point of skeletal joints. Such a musculoskeletal description makes it possible to investigate the semantic processes that arise from elementary gestural phenomenon and that would otherwise be difficult to distinguish. This approach involves a radical shift of perspective that has profound consequences in the way the different gestural components are perceived and represented.

The representation framework of TYPANNOT is based on the kinesiologic model presented in section (§ 3.2). It distinguishes the different articulatory domains that provide intrinsic representations for the five parameters that structure SL. Four are static articulatory parameters (HS, LOCini, MouthAction and EyesAction[5]) and one is a dynamic parameter (MOV) that describes the way an articulatory parameter is being transformed. This representation framework must also meet the practical aspects of transcription which implies processing information in the form of viewable, transferable and searchable textual data. In order to translate this representation framework into a viable typographic system, four requirements have been identified, which guide the design process: *genericity*, *readability*, *modularity* and *inscribability*.

2.2.   Design principles

2.2.1.   Genericity

The first requirement directly stems from the phonological transcription approach using an articulatory model of the human body. For each of the articulatory parameters, gesture is deconstructed into discrete elements representing four layers of information (Fig. 6):
 – Layer 1: the SL parameter that the transcription refers to (e.g., handshape);
 – Layer 2: the different parts that compose the parameter (e.g., thumb);
 – Layer 3: the different variables associated with each part (e.g., angle);
 – Layer 4: the values assigned to those variables (e.g., open).

Each layer has a limited set of characteristics that defines it, creating individual bricks of information. Once defined, these characteristics form the generic components of the TYPANNOT transcription system. Symbolic graphic representations can be assigned to them and later encoded into a font to perform like letters.

––––––––––––––––––––

5. MouthAction and EyesAction are two parameters for describing the posture of the mouth and that of eyes and nose.

FIGURE 6. This transcription of a mouth action has been set in generic form and colorized in order to distinguish the four layers of information. The SL parameter (layer 1) is written in black. The parts (layer 2) are written in orange. The variables (layer 3) are written in green. The values (layer 4) are written in blue.

While there might be hundreds of thousands of possible configurations for a parameter (261 million possible configurations for the handshape alone), TYPANNOT requires only a few generic components to describe them all. The systematic organization of the information into four layers also gives the transcription a robust syntax that ensures it can be consistently produced, manipulated, and searched. Finally, through the principle of genericity, TYPANNOT allows annotators to perform queries and comparisons throughout different phonological levels, involving a combination of features or targeting a single one. This kind of deep querying of data is impossible to perform with other SL representation systems.

2.2.2. *Readability*

The TYPANNOT phonological approach aims to provide a discrete and low-level representation of gestures. From a typographical point of view, this is achieved at the cost of linearity. While the generic design principle involves methodically decomposing gestures into a suite of individualized pieces of information, it breaks down the only visuo-spatial guiding perspective that would allow users to read gestures in an intuitive and instantaneous way: the body space, in other words a unified representation of the body. For a language that is fundamentally visual in terms of perception, it is ironic that its representation needs to distill it to its lowest distinctive components, thus making its transcription unreadable. Although logographic systems like SW exist and show how readability in SL can be achieved through a spatialized representation

of the parametric components, none are able to retain a high level of discreteness while doing so.

This trade-off limits the main function of a transcription system: analysis. To be relevant to the principles of both genericity and readability, we believe that a SL transcription system needs to be able to display the same information in two formats: 1) a *generic* form that shows the distinct bricks of information organized through robust linear syntax that allows deep research into the gesture components; 2) a *composed* form that translates and integrates the different phonological components into a recognizable form: the image of the signed body. Progress in font encoding technologies (i.e., OpenType[6] features) and typographic functionalities (i.e., contextual ligatures) allows us to design a system that gives users the ability to seamlessly display one form or the other while retaining data integrity. For each of the articulatory parameters, we define a specific graphic formula that translates the generic pieces of information into a unified and visually explicit "composed" glyph. For example, the initial location (LOCini) parameter refers to the structure of the upper limbs. This structure is made out of three parts (arm, forearm, hand) that are articulated according to various variables (flexion/extension, adduction/abduction, etc.) and their possible values (neutral, +1, +2, etc.). Displayed in the *generic* form (Fig. 7), the transcription looks like a string of symbols following a linear syntax. Displayed in the *composed* form (Fig. 8), the transcription looks like an articulated structure with joints (shoulders, elbows, wrists) and segments (forearms) forming an expressive figure with the head (triangle). The last segment, the hand, is shown on each side in order to appear bigger and more readable.
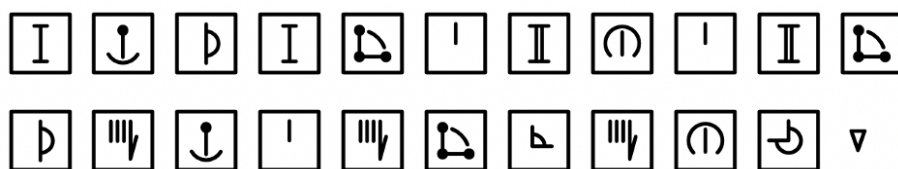
FIGURE 7. LOCini displayed in the generic form (left side only)

---

6. OpenType is a vectorial font format that allows encoding any character associated with Unicode, regardless of the platform (Mac, Windows, Android, etc.); OpenType fonts can have advanced typographic features that allow handling complex writing and typographic effects like ligatures.

FIGURE 8. LOCini displayed in the composed form (left and right sides)

While the logographic composed form has no analytic function, it reflects the ongoing ethical commitment of our team to provide accessible representation tools for both linguists and signers.

2.2.3. *Modularity*

Designing a typographic system that is both phonological and logographical means that we have to maintain strict equivalence between the two forms. This equivalence can be achieved through using a modular design approach. *Composed* forms are basically projections of intrinsic articulatory characteristics following an allocentric perspective[7]. Graphic modules symbolizing the different parts (i.e., fingers and thumb) of an articulatory system (i.e., the hand) are transformed and assembled according to articulatory characteristics (i.e., form, angle, contact, etc.) inside a framework that systematically replicates the spatial organization of the body (Fig. 9).

FIGURE 9. Three composed HS glyphs showing variation in their construction

This modular design principle helps us solve the question of equivalence, and more importantly, allows us to automatize the glyph creation process. To this end, the articulatory approach is synonymous with massive combinatorial possibilities and we are now facing the problem of quantity. For example, the articulatory system of the hand alone can give rise to hundreds of thousands of configurations and thus requires the production of equal amounts of composed glyphs. Manually design-

---

7. In an allocentric perspective, the position of a body part is defined relative to the position of other body parts (e.g., the position of the hand depends on the position relative to the forearm, which depends on the position relative to the arm). In an egocentric perspective, the position of an element depends on the orientation of the viewer's body (e.g., the hand is in front, on the left and up).

ing that many glyphs is, for many reasons, impractical[8]. Thanks to our modular framework and scriptable font design environment (i.e., Robofont[9]), we are able to code the module's integration process in order to generate all the composed forms.

2.2.4. Inscribability

The TYPANNOT project aims at providing a tool for the representation of SLs, but also explores the possible relations between the annotator or signer and a written form. While vocal writing systems use conventional graphic principles, SL writing or transcription systems have the unique opportunity to engage in a dialogue between writers and the constitutive dimensions of their own language: the human body. Through the intrinsic perspective of the articulatory models and a typographic system that combines phonographic and logographic dimensions, GestualScript believes that annotators can develop an intuitive bond with their transcription, not by describing "what SL looks like" but by recognizing "how SL happens" and describing it from the inside. Because it is not a familiar way to perceive gesture, this shift in perspective involves the creation of specific input interfaces that promote interactions that elucidate the articulatory and dynamic principles behind it (gesture). While designing those interactive input interfaces (see § 2.4) we are aiming at facilitating the process of incorporating and assigning the transcription systems (Poizat, Salini, and Durand, 2013).

2.3. Corpus and User-Driven Font Systems

OpenType fonts can contain up to 65,536 glyphs. With TYPANNOT, following the principle of modularity (§ 2.2.3), generating all morphologically possible combinations of a parameter's elements greatly exceeds the maximum capacity of font glyphs. For example, the automatic composition of TYPANNOT handshapes creates 291,600 glyphs.

To create TYPANNOT fonts, GestualScript has to decide what criteria should be used to reduce the number of possibilities by selecting the

---

8. One reason among others: at the improbable rate of one compounded glyph per second, it would take 261,000,000 seconds to encode all handshapes, which would require working more than 8 years, 24/7.

9. Robofont is software for typeface creation, which can automatically generate contextual ligatures from graphic modules and instructions on their layout. For the development of TYPANNOT, Frederik Berlaen, developer of Robofont (`https://robofont.com`), has kindly provided GestualScript with a license to use his software.

most appropriate and relevant glyphs, keeping in mind language evolution. A bottom-up approach[10] was chosen as the operating principle:

- First, a character set was created using the 237 handshapes identified by Eccarius and Brentari (2008) in their corpus made of confirmed configurations present in lexicons of 9 SLs (Hong Kong, Japanese, British, Swedish, Israeli, Danish, German, SwissGerman and American SL).
- A further development consisted in extending the character set to include a larger sample of signs, thanks to the addition of the configurations listed in the inventory of SW (Sutton, 1995), plus some variants sought for completeness.
- A third, future step will expand the character set in a participative way. By using TYPANNOT, linguists from every background will transcribe handshapes that haven't got yet a composed form. The program will automatically identify and collect those unknown forms in order to plan regular updates of the character set (§ 2.4.3).

This updating procedure will also help us identify and register new handshapes in SLs that are less studied, expending their understanding.

2.4. Input Systems

Setting up a complete and comprehensible typographic system for SLs was no easy task. The methods described above were essential in the completion of TYPANNOT's goal to offer an efficient solution to enhance linguistic research on SL. Yet, the typographic system by itself is not sufficient. In order to truly come into being, a custom tool that allows researchers to use TYPANNOT to its fullest capacities was needed. Therefore, we are currently shaping the TYPANNOT Keyboard into a digital interface, which will offer several input devices to fit a wide spectrum of transcription approaches.

2.4.1. *Enhancing Knowledge Through Technology and Design Efficiency*

Creating a digital interface to make TYPANNOT fully accessible goes far beyond the sole possibility of combining glyphs together in order to inject them into office software (e.g., Word, PowerPoint, etc.) or multimodal transcription software (e.g., ELAN).

Such an interface has the responsibility of ensuring that users will understand and use TYPANNOT in a coherent and consistent manner. It is true that a well thought out user experience is always essential for

---

10. The selection is not based on given linguistic rules but on the occurrences found in various linguistic corpus.

any given tool. But in this case, it goes beyond the necessity of user-friendly software. TYPANNOT, as a new typographic system, needs to be discovered, understood and used in a consistent way. It is an essential part of the process to ensure that transcriptions using TYPANNOT can be understood and used in cross referencing research.

That means that the structure and the interface design have to be engineered to give users key pieces of information about TYPANNOT itself: information on the structure of the typographic system, on the value of each glyph, on how to combine them properly, and what the results signify.

This task can be achieved in various ways. Some are very tangible, like a quick overview of the software or a series of tutorial videos and exercises to display the full potential of the interface. Others, equally important, are less tangible, like the overall interface design and inter-active feedback to help users understand what they can do and how to do it (Fig. 10).
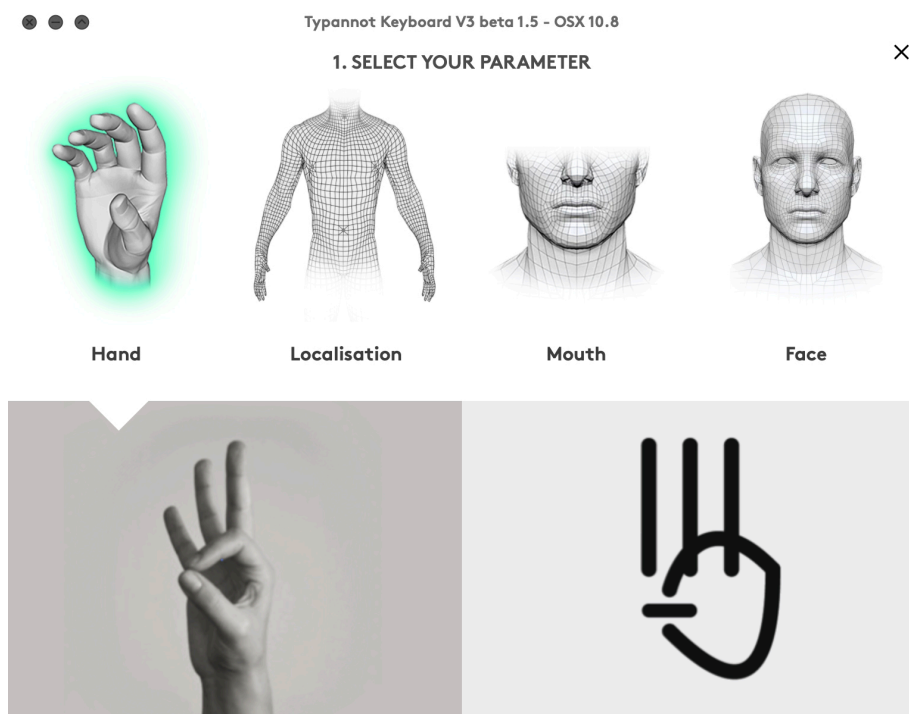
FIGURE 10. Home page of TYPANNOT Keyboard: the interface design guides the user on what can be done and how it should be done

2.4.2.   *Opening Possibilities for Present and Future Research*

To be efficient, the design work of the TYPANNOT Keyboard has to take into account the TYPANNOT typographic system and different user profiles. Our opinion is that the latter is the most important aspect. Understanding the user's thought process and motivation is essential. And this alone is a substantial task. The TYPANNOT Keyboard is not intended to shape the direction of researchers' work. Our goal is to offer a flexible tool that enhances research capabilities, while retaining as much information as possible to broaden our understanding of SL. To this end, the TYPANNOT Keyboard and its interfaces are being developed to fit all research methodologies, all types of focus and specializations, and of course, all SLs and gestural actions.

From the beginning, the TYPANNOT Keyboard has been designed to be a virtual keyboard that can be used on top of any given software. Two different interfaces were developed, each offering its own transcription experience. These interfaces have three main features in common: an interactive 3-dimensional representation of the parameter, the corresponding glyph in TYPANNOT Font, and an input device.

The Parametric Interface (Fig. 11) has TYPANNOT's 4 layers of information (parameter, part, variable, value) as the input device. Glyphs are composed by selecting and adding values. It is a very simple way to compose glyphs that ensure a perfect comprehension of the typographic system.

The Gestural Interface (Fig. 12) uses motion capture devices (Leap Motion, Neuron Perception, Brekel Pro Face 2) to offer an effortless transcription process. This means that the annotator's own body is used to transcribe, directly reproducing the handshape, body position or facial action. This offers a very intuitive input system that truly connects with the nature of SL.

2.4.3.   *Research Sourced Typographic Library*

Beyond a learning tool and an input device, the TYPANNOT Keyboard is also the answer to the technological limitations of OpenType fonts. In its initial version, the keyboard will be loaded with 990 glyphs corresponding to our fundamental set. But 990 is not exhaustive, and researchers will inevitably need more glyphs. When users compose a glyph needed for their own research and not yet included in the glyphic library, the TYPANNOT Keyboard will offer them the opportunity to request the addition of the new glyph. On a regular basis, the TYPANNOT library will be updated, including all requested glyphs. With this open sourced process, TYPANNOT will be the research-based font that includes all glyphs from all research around the globe.
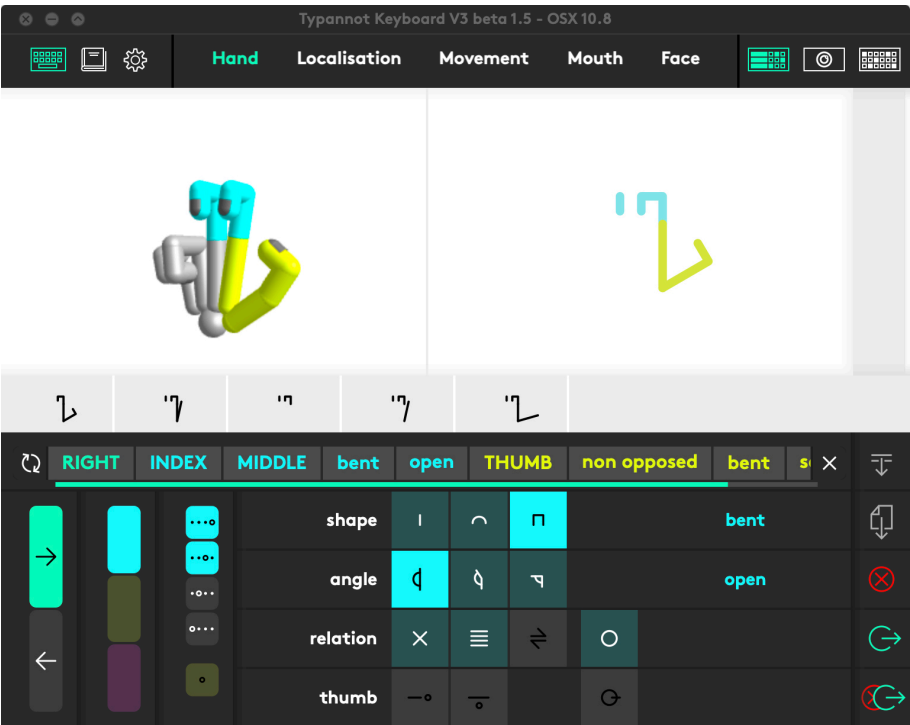
FIGURE 11. Parametric Interface of the TYPANNOT Keyboard for HS Coding

3. The Kinesiological Approach

3.1. The Body at the Center of Linguistic Analysis

It is further important to recognize the role of movement in the expression of meaning in both SL and gesture, and therefore, to recognize the importance of movement in the development of our research. While phonological studies endeavor to faithfully represent the other manual and non-manual components in order to analyze them, any attempt to fix movement seems to go against the very essence of its ephemeral nature. The strong physical anchoring of movements generates a great complexity of representation; this creates analytical difficulties, which contribute to the marginalization of movement research relating to research about SLs and gestures in all types of communication. In turn, this results in a poor understanding of the nature and meaning of move-
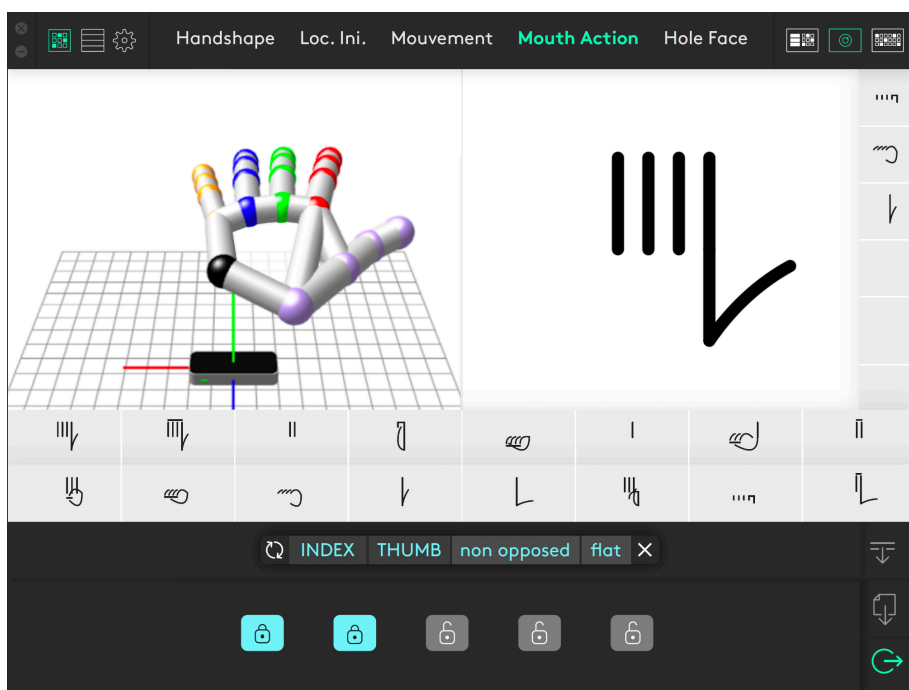
FIGURE 12. Gestural interface of the TYPANNOT Keyboard for HS coding

ment, all of which ends up reinforcing the complexity of movement representation: it is a vicious circle[11].

Besides being marginalized, the study of movement that does exist is almost exclusively focused on the activities of the hand, whose capacity to produce meaning is never in doubt. But while researchers are increasingly interested in the participation of the face and the trunk in the production of meaning, the arm and forearm remain confined to the role of simple connecting segments.

Breaking the ruts created by these common trends—i.e., the focus on the hand and the marginalization of the movement—requires a radical change of approach, which is precisely what Dominique Boutet[12] proposes through the kinesiological approach developed in his own re-

---

11. Getting out of this loop is a very topical issue, practical but also theoretical. Indeed, the deepening of the analysis of movement is perceived as a possible response to the debate which animates research in SL on the distinction between co-verbal gestures and purely linguistic phenomena (Goldin-Meadow and Brentari, 2017).

12. Dominique Boutet was coordinator of the GestualScript team from its beginnings in 2008 until 2020, when he succumbed to the COVID-19 pandemic. Parallel to his commitment to the representation of SLs, he developed the kinesiological ap-

search (Boutet, 2018). Taking into account the physical and physiological mechanisms governing the anatomical constraints of human movements, Boutet seeks both to restore the body capacities as a vehicle of meaning (see § 3.2) and to show that it is possible to describe (and analyze) the movement faithfully and efficiently. For this, it is necessary to make innovative choices: include all of the upper limbs, change the frame of reference, and abandon Euclidean geometry.

3.1.1.   *Considering the Upper Limb*

The hand is often considered as the only articulator carrying movement and, consequently, meaning (see Intro and § 1). However, this is not an entity independent of the rest of the body. It is attached to the forearm, which is attached to the arm, itself linked to the trunk. The hand is therefore the most distal[13] end of a chain of segments (SEG) comprising the forearm as well as the arm, and this concatenation necessarily generates a series of physiological constraints and limitations on the freedom of movement of each of these SEGs.

In the approach proposed by Boutet, movement is carried by all the SEGs of the upper limb, considered as an articulatory system. The movements and postures of each of these SEGs are described according to principles governed by biomechanics (Kapandji, 1997). Each SEG is associated with so-called degrees of freedom (DoF), which correspond to the rotation of a SEG around an axis located at the level of the proximal adjacent SEG. Thus, the hand will be described in relation to its position relative to the forearm; the latter will be linked to the arm, which in turn will be described in relation to the trunk. These axes mainly pass through the joints (wrist, elbow, shoulder), but they can also cross bones longitudinally (ulna + radius, humerus).

The upper limb is therefore an "infrastructure which underlies all the possible movements" (Boutet, 2010, p. 2) of the hand, and constitutes an articulated whole with inseparable parts, all having a precise role in the unfolding of the sign.

3.1.2.   *Changing the Frame of Reference (FoR)*

An articulatory approach has a profound effect on how movement is inscribed in the representation space. Traditionally, the description of

---

proach to human gestures, which has greatly influenced the ongoing work carried out by the GestualScript team.

13. Distal and proximal are concepts indicating the position of a Seg relative to another Seg and to the body: a Seg is distal if it is located further from the body in relation to another Seg (the hand is distal in relation to the forearm which is distal to the arm); a Seg is proximal when closer to the body than another Seg (the arm is proximal to the forearm which is proximal to the hand).

movement is done in a space defined by Cartesian planes: horizontal, vertical and sagittal. The point where these planes intersect defines the frame of reference (FOR), which can be absolute, relative and intrinsic (Levinson, 1996; see Fig. 13). The first is centered on the surrounding space (the choice of a specific location on the space giving rise to several subtypes of relative FORs); the second on the body of the signer (again, different subtypes are possible); the third is centered on an object and is defined on the basis of its inherent characteristics.

Absolute FoR
« The hand is on the west side »

Relative FoR
« The hand is left of the tree »

Intrinsic FoR
« The hand is on the right of the signer's shoulder »
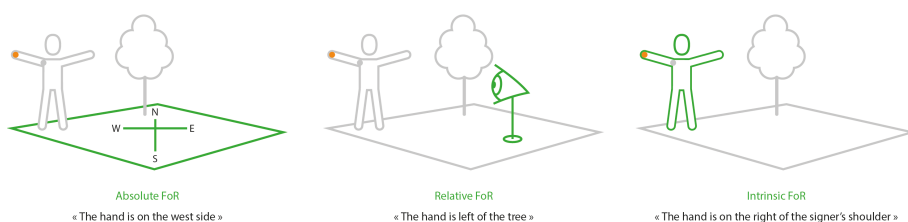
FIGURE 13. Absolute, relative or intrinsic frame of reference (Levinson, 1996)

The analysis of the SL representation systems (see §1) shows that they adopt FORs which can be relative or absolute (never intrinsic), but also that within the same system, different FORs can be adopted, sometimes on a case-by-case basis.[14]. In these events, these descriptive instabilities give rise to fragmentations of the description space even if looking just at the movement, thus generating the risk of inconsistencies in the analysis of signs (see above §1).

Extending the analysis to all the SEGs of the upper limb, the kinesiological approach risks being confronted with a multiplication of the difficulties of representation and analysis, unless it adopts a coherent system of registration in a single typology of FOR. The choice has been to abandon the projection of SEGs on planes in favor of a parameterization of the SEGs in their own respective space. This is allowed by the use of intrinsic FORs. The description of each SEG is then centered on an object (i.e., the proximal SEG adjacent to the analyzed SEG) and is defined on the basis of the intrinsic characteristics of this same object (which are in fact equivalent to the DoF of the analyzed SEG). The various DoFs are identified by the name of their poles (or joint stops): abduction (ABD) on one side and adduction (ADD) on the other, flexion (FLX) and extension (EXT), pronation (PRO) and supination (SUP), internal rotation (RIN) and external rotation (REX).

---

14. In her thesis, Bianchini (2012) offers a detailed analysis of all the FORs present in SW, showing that the FOR can vary within the same parameter.

More concretely (see Fig. 14), the hand—whose FoR is defined in relation to the forearm—is affected by 2 DoF whose poles are Flx-Ext and Abd-Add, both passing through the wrist joint; a 3rd DoF is present, Pro-Sup, which goes through the ulna and the radius (bone that can cross). This latter DoF could be considered to affect the forearm, but since the "result" of this movement is visible on the hand, it was decided to include it in the description of the hand. The forearm—more proximal than the hand and more distal than the arm—is affected by 2 DoF, i.e., Flx-Ext which passes through the elbow, and Rin-Rex which is due to the possibility of the head of the humerus to rotate in the scapula. Here too, although Rin-Rex is located on the upper arm, it is assigned to the forearm because its result is visible there. Finally, the arm—the most proximal of the Segs and which is described in relation to the trunk—is affected by 2 DoF, i.e., Flx-Ext and Abd-Add, both passing through the shoulder joint.

Figure 14. Listing of the degrees of freedom (DoF) of the segments (Seg) of the upper limb

The FoR used by the kinesiological approach is therefore unique, because in fact it is an intrinsic FoR, but it is combinatorial too, since it takes into account the fact that in an articulatory system each Seg depends on its proximal Seg. This innovative choice has also the advantage of being ready for the envisaged technological requirements: motion capture systems (MoCap) are gaining in importance in gestural

studies and some of these technologies are based on intrinsic FoRs, but the classical representation systems used to analyze them are still based on relative or absolute FoRs, thus requiring a conversion. The kinesiological approach allows direct access to data, minimizing biases related to said FoR conversion, thus facilitating not only the understanding of the movement but also its representation: this should make it possible to break the vicious circle discussed in §3.2.

The change of FoR makes it possible to focus on the possibilities and limits of the movements specific to each of the SEGs, but it also generates other modifications: bypassing Cartesian planes for the description of SEGs requires finding a geometry that can take into account elements which no longer fit into these plans.

3.1.3.   *The Transition to a Non-Euclidean Geometry*

In 1934, Ernest A. Codman affirms that if the arm is completely raised, it is both in complete Rex and complete Rin (Codman, 1934). Neither its author nor the specialists in movement and physiology who looked into it (Pearl et al., 1992) succeeded in explaining this fact. It is ultimately the abandonment of Euclidean geometry in favor of a non-Euclidean geometry which allows understanding the existence of diadocal movements[15] (MacConaill, 1953) and therefore resolving this alleged paradox.

Euclidean geometry draws its forms on planes and is based on 5 postulates, the last of which states (simplifying) that "given a straight-line *d* and a point *P* located outside it, there is one and only one straight line *d'* passing through *P* and parallel to *d* (Fig. 15). The non-Euclidean geometry proposed by Gauss, a.k.a. spherical geometry, rejects this postulate, asserting that "there exists an infinity of lines passing through *P* which are parallel to *d* (Fig. 16). This is possible if, and only if, we abandon the representation on the plane in favor of a representation on the sphere. A line (spherical) will be a circle drawn at the "equator" of a sphere; a point will be a pole where several spherical lines intersect. The consequence is such that—unlike the principles stated by Euclidean geometry—in spherical geometry, the curve (or spherical line) is a very simple plot and the plane straight line is a complex figure.

Coming back to the description of the body, the most proximal end of a SEG (e.g., the elbow for the forearm) constitutes the center of a (portion of) sphere; the movement of the different DoFs draw at the most distal end of the SEG (e.g., the wrist for the forearm) "straight" spherical

---

15. The Codman's Paradox is the result of a "diadocal movement," i.e., an involuntary movement which, on a SEG with 3 DoF (like the hand or the arm, if considering also the hidden DoF carried by the humerus), affects a DoF when the other two moves consecutively.

Postulate 1
Between any two points, forming a segment AB, there exists a straight line, d.

Postulate 2
Unlike a segment, which is finite at outer points A and B, a line is infinitely extendable at both ends, d.

Postulate 3
A circle is a series of points, P, equally distant from one central point, O.

Postulate 4
All right angles are congruent.

Postulate 5
Two lines, which are drawn with a third intersecting in such a way that the intersecting angles are two right angles, meaning they are parallel, can be extended infinitely without intersecting each other.

FIGURE 15. The postulates of Euclidean geometry

Postulate 1
Between any two points, forming a segment AB, there exists a straight line, d.

Postulate 2
Unlike a segment, which is finite at outer points A and B, a line is infinitely extendable at both ends, d.

Postulate 3
A circle is a series of points, P, equally distant from one central point, O.

Postulate 4
All right angles are congruent.

Postulate 5
When drawn on a sphere, Euclid's parallel postulate does not stand. For example, two lines that are drawn with a third that intersects at a right angle may intersect the perpendicular line twice, and may not themselves be extended infinitely.

FIGURE 16. In spherical geometry, Euclid's 5th postulate is not respected

lines which intersect at the poles of this sphere. The use of the spherical geometry, associated with the multiple intrinsic FoRs, thus allows to describe in a simple way what is anatomically simple—that is to say the production of curves—and in a complex way what is complex for the body, i.e., the straight lines (Fig. 17). This therefore contributes to the creation of a faithful and efficient description of the movement.

3.2. The Body as a Generator of Meaning

In the classical approach to SL analysis, the hand is seen as the articulator which, replacing the mouth, conveys meaning. The kinesiological approach, with its consideration of the entire upper limb in a non-
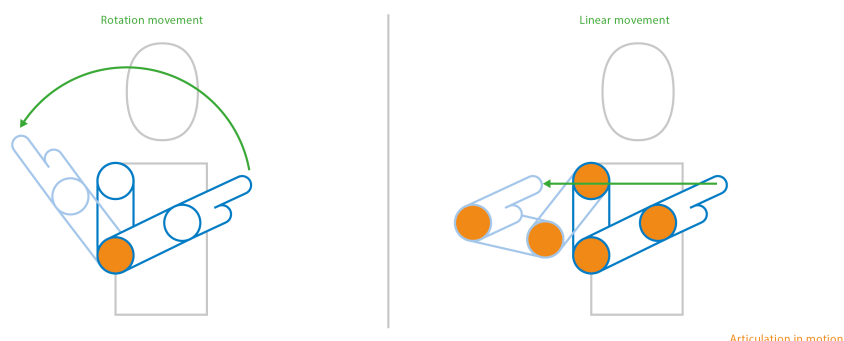
FIGURE 17. Segment(s) performing a simple (1 DoF) or complex (several DoFs) anatomical gesture

Euclidean geometry and with intrinsic and multiple FoRs, questions the validity of this idea.

For example, the co-verbal gesture "no" can be done, at least in France, by standing with the arm alongside the body, the forearm slightly bent, and the hand extended; it is then possible to perform repeated movements of ABD and ADD of the hand (a "little no"). But to support the disagreement, it will then not only be the hand that will be in motion, but also the forearm, and why not, the arm (a "big no"). Visually, these two realizations are very different, but nobody doubts that they convey the same meaning "no": how is that possible?

To answer this question, the kinesiological approach proposes to search for the structural invariants of the articulatory dynamics which underlie the creation of signs and which are hidden by purely visual differences. Once again, innovative choices and new concepts become necessary: a) proposing a new notion of movement and temporality; b) restructuring the classic parameters of SL analysis.

3.2.1.   *Movements and Temporality*

The search for invariants begins with understanding the different typologies of movement. Boutet suggests distinguishing proper movement from displacements and transfers. It is a question of proper movement when a SEG initiates movement, that is, at least one of its DoFs performs a rotation. If this SEG has SEGs more distal than itself, these—driven by the proper movement—displace in space, without even their DoF moving. Finally, in special cases, there may be an inertial transfer of movement: the rotation of a DoF then engages the variation of a DoF on a different SEG.

If a movement can propagate between different SEGs, then it is possible to study its flow (Boutet, 2018), that is to say the order in which the different SEGs are set in motion. If a movement begins on the hand and then continues on the forearm, this is a distal-proximal flow; if the reverse is true, then it is a proximal-distal flow; the flow can also be "neutral" if all the concerned SEGs move at the same time, or even "absent".

To come back to the "big no" and "small no," they are identified as manifestations of the same sign because they correspond to the same pattern: the hand initiates a repeated movement on the DoF ABD-ADD, which is propagated by following a distal-proximal flow. This reality, however, remains hidden under the superficial manifestation of these two demonstrations. A restructuring of the classic manual parameters (see Intro) is therefore necessary.

3.2.2.   Restructuring of Location and Orientation

Boutet (ibid.) argues that the search for invariants can be facilitated by restructuring the classic parameters of orientation, location and movement, proposing to replace them by parameters whose names may seem similar, but whose scope will be radically different: the initial location (LOCini) and the movement (Mov).

The LOCini makes it possible to fix the position of all the SEGs before the deployment of the sign. Therefore, it brings together the notions of orientation and location, but by extending them to the whole of the upper limb. Concretely, the LOCini is described through the angles of rotation (in an intrinsic FoR) of all the DoFs of the SEGs (only 7 in total).

The kinesiological approach then makes the hypothesis (Chevrefils, forthcoming) that once the body is installed in a posture, the resulting Mov is simple: the body's tendency to decrease the DoFs to be controlled pushes the SEGs to coordinate (Turvey, 1990) and to prefer a distal-proximal flow,[16] which leads to economy and predictability of movements. The results of a first study involving a few minutes of corpus in three SLs (English, French and Italian) seem promising (Danet et al., 2017). A deepening of this hypothesis, through an accurate analysis of the kinematic data from a MoCap system, is underway.

Therefore, the subdivision between LOCini and Mov also contributes to understanding the difference between the small and the big

---

16. Despite the existence of a decreasing inertial slope from the arm to the hand (Dumas, Chèze, and Verriest, 2007) favoring a proximal-distal flow, the communicational aim of movement in SL would reverse this trend: the flow of movement would then be predominantly distal-proximal (Chevrefils, forthcoming). The preliminary study of a corpus of three SLs seems to confirm this trend (Danet et al., 2017), the causes of which are still under investigation.

"no": the two are identified as distinct realizations of the same sign because their Mov is the same, despite a difference in the LocINI and in the Mov amplitude. A difference in the SEG initiating the gesture, in the DoF concerned or in the flow would not have allowed these signs to be identified as "no". The kinesiological approach thus renders to the whole body (and no longer just to the hand) its function of generator of meaning.

3.3.   From Theory to Practice: From Movement to TYPANNOT

Although offering a reliable and economical description of the movement, the kinesiological approach requires the handling of many concepts. This novelty could generate the impression that this description system can only be used after following a specific theoretical training, in particular concerning the use of intrinsic FoR which is "experienced" by any speaker, but which is not "recognized" and "perceived" by most of them.

The work of the GestualScript team addresses this point. Its goal is to make all these notions accessible and functional through the creation of TYPANNOT, a transcription system based on typographic principles. TYPANNOT is not only a graphic formalization of a theory, it is the instrument for appropriating the theory itself, drawing its bases from the kinesiological model and its descriptive efficiency, it will allows the constitution of a readable, writable, and searchable corpus (of SL or of co-verbal gestures) readable, scriptable and searchable according to the desired level of granularity.

The passage from the complexity of a theoretical approach to the intuitiveness of a "turnkey" typographic system requires answering several preliminary questions, including a non-exhaustive list of which is: how can such complex phonological descriptions be readable and scriptable? how to increase the descriptive precision of the system without increasing the transcription time, or even reducing it? how to make the transcriptor conscious of their own body so that the notion of intrinsic FoR is understandable?

The answers to these questions go through the definition of different layers of information and construction principles, set out above.

Discussion

TYPANNOT is a typographic writing system intended for linguists needing to study and transcribe SLs. It follows a musculoskeletal articulatory approach that changes the conventional perspective from which

gestures are observed. This perspective allows researchers to investigate how corporal activity fully determines the construction, modulation, and transformation of meaning of the signs in SLs.

Further, the phonological articulatory approach of TYPANNOT, in which the joints of the skeleton are represented in order to distinguish the abstract phonetic units that correspond to meaning in SLs and movement, it possible to transcribe gestures using the same corporal parameters. This way, TYPANNOT is also a tool for investigating other forms of corporal expression, such as co-verbal gesturality, which refer to gestures made while talking.

Also, because our writing system describes the morphological components of SLs at the articulatory level, TYPANNOT indexes SL using elementary characters. This makes it possible to refer to SL signs using simple morphological and gestural features rather than translating them into a vocal language, as is systematically the case with online SL dictionaries.

Another aspect of the TYPANNOT system is its ability to transcribe the dynamics of the articulatory system. Such transcription possibilities can be coupled with motion capture technologies to explore new ways of inputting, recognizing, and reproducing SL signs. When a stream of recorded gestures can be directly recognized to automatically generate a transcription (input), this transcription can also generate the 3D animation of a signing avatar (output).

Finally, although not the main goal of this project, TYPANNOT is fitting to the contribution of the development of a writing system for SLs by giving signers a new form of expression that is based on the human body itself, the very center and origin of all SL expression.

References

Augustus, Robert A., Elsie Ritchie, and Suzanne Stecker (2013). *The Official American Sign Language Writing Textbook*. si5s.org & ASLized.org.

Bébian, Auguste (1825). *Mimographie ou essai d'écriture mimique, propre à régulariser le langage des sourds-muets*. Paris: Louis Colas.

Bianchini, Claudia S. (2012). "Analyse métalinguistique de l'émergence d'un système d'écriture des langues des signes: SignWriting et son application à la langue des signes italienne (LIS)." PhD thesis. Université de Paris 8 Vincennes & Università degli Studi di Perugia.

———— (2016). "Regard sur la nature de SignWriting (SW), un système pour représenter les langues des signes (LS)." In: *Dossiers d'HEL*. Vol. 9, pp. 404–421.

Boutet, Dominique (2010). "Structuration physiologique de la gestuelle: modèle et tests." In: *Revue de Linguistique et de Didactique des Langue* 42, pp. 77–96.

Boutet, Dominique (2018). "Pour une approche kinésiologique de la gestualité: synthèse." Habilitation à Diriger des Recherches. Université de Rouen Normandie.

Braffort, Annelies (1996). "Reconnaissance et compréhension de gestes, application à la langue des signes." PhD thesis. Université de Paris 11 Orsay.

Brugeille, Jean-Louis (2007). "L. S. F. numérique." In: *Journée professionnalisante sur la traduction—traduction et métiers émergents: traducteur en langue des signes. Université de Toulouse Le Mirail*.

Chevrefils, Léa (forthcoming). "Vers une modélisation des constituants gestuels des signes: capture de mouvement et transcription formelle d'un corpus de langue des signes française." PhD thesis. Université de Rouen-Normandie.

Codman, Ernest (1934). *The Shoulder: Rupture of the Supraspinatus Tendon and other Lesions in or about the Subacromial Bursa*. Boston: R. E. Kreiger.

Cuxac, Christian (2001). "Les langues des signes: analyseurs de la faculté de langage." In: *AILE—Acquisition et iteraction en langue étrangère* 15, pp. 11–36.

Danet, Claire et al. (2017). "Structural Correlation Between Location and Movement of Signs: Lacking Motion Economy for Co-Speech Gestures." In: *Language as a Form of Action*. Rome: CNR.

Dumas, Raphaël, Laurence Chèze, and Jean-Pierre Verriest (2007). "Adjustments to "McConville *et al.* and Young *et al.* Body Segment Inertial Parameters"." In: *Journal of Biomechanics* 40.3, pp. 543–553.

Eccarius, Petra and Diane Brentari (2008). "Handshape Coding Made Easier: a Theoretically Based Notation for Phonological Transcription." In: *Sign Language & Linguistics* 11.1, pp. 69–101.

Goldin-Meadow, Susan and Diane Brentari (2017). "Gesture, Sign, and Language: the Coming of Age of Sign Language and Gesture Studies." In: *Behavioral and brain sciences*. Cambridge: University Press, pp. 1–60.

Guitteny, Pierre (2007). "Langue des Signes et Schémas." In: *TAL* 48.3, pp. 1–25.

Hanke, Thomas (2004). "HamNoSys: Representing Sign Language Data in Language Resources and Language Processing Contexts." In: *Workshop on the Representation and Processing of Sign Languages on the Occasion of the Fourth International Conference on Language Resources and Evaluation (LREC Lisbonne)*, pp. 1–6.

Haouam-Bourgeois, Nadia (2007). *La Signographie: Master 1*. Les Essart-le-Roi: Éditions du Fox.

Jouison, Paul (1995). *Écrits sur la Langue des Signes Française (LSF)*. édition critique établie par Brigitte Garcia. Paris: L'Harmattan.

Kapandji, Ibrahim-Adalbert (1997). *Physiologie articulaire. 1, Membre supérieur*. Paris: Maloine.

Levinson, Stephen C. (1996). "Frames of Reference and Molyneux's Question: Crosslinguistic Evidence." In: *Language, Speech, and Commu-*

nication: Language and Space. Ed. by P. Bloom et al. Cambridge, MS: The MIT Press, pp. 109–169.

MacConaill, Michael A. (1953). "Movements of Bone and Joints: 5—The Significance of Shape." In: *The Journal of Bone and Joint Surgery* 35.2, pp. 290–297. DOI: doi:10.1302/0301-620X.35B2.290.

Pearl, Michael L. et al. (1992). "Codman's Paradox: Sixty Years Later." In: *Journal of Shoulder & Elbow Surgery* 1.4, pp. 219–225. DOI: doi:10.1016/1058-2746(92)90017-W.

Poizat, Germain, Deli Salini, and Marc Durand (2013). "Approche énactive de l'activité humaine, simplexité et conception de formations professionnelles." In: *Education, Sciences & Society* 4.1, pp. 97–112.

Prillwitz, Siegmund et al. (1989). *Hamburg Notation System for Sign Languages: an Introductory Guide*. Hamburg: Signum Press.

Sandler, Wendy and Diane Lillo-Martin (2006). *Sign Language and Linguistic Universals*. Cambridge: Cambridge University Press.

Slobin, Dan I. et al. (2001). "Sign Language Transcription at the Level of Meaning Components: the Berkeley Transcription System (BTS)." In: *Sign Language & Linguistics* 4.1–2, pp. 63–104.

Stokoe, William (1960). "Sign Language Structure: an Outline of the Visual Communication Systems of the American Deaf." In: *Journal of Deaf Studies and Deaf Education* 10, pp. 3–34.

Sutton, Valerie (1995). *Lessons in SignWriting*. La Jolla, CA: DAC.

———— (2020). "SignWriting History. SignWriting Official Website." URL: http://www.signwriting.org/library/history/index.html (visited on (9/20)).

Theureau, Jacques (2004). *Le cours d'action: méthode élémentaire*. Toulouse: Octares.

Turvey, Michael T. (1990). "Coordination." In: *American Psychologist* 45, pp. 938–953.

Varela, Francisco, Evan Thompson, and Eleanor Rosch (1993). *L'inscription corporelle de l'esprit: sciences cognitives et expérience humaine*. Paris: Éditions du Seuil.

How to Improve Metalinguistic Awareness by Writing a Language Without Writing:
Sign Languages and Signwriting

Claudia S. Bianchini

Abstract. Multilingualism permits to compare elements of each known language, favoring the development of metalinguistic awareness which helps to correlate the functioning of own reference languages (L1, L2, etc.). To all intents and purposes, signing deaf are bilingual people with sign language (SL) as L1 and an own vocal language (VL) as L2. Since deafness affects only the auditory canal, it should be reasonable to expect that deaf (signing or not) have the same competence in VL writing than hearing people; however, Gillot (1998) has demonstrated that 80% of French deaf adults have a scarce level of literacy. On the other hand, Garcia et al. (2007) have proved that deaf signers have a better relationship with writing than deaf not knowing SL, while Perini (2013) has evinced how SL knowledge helps the deaf to understand the functions of writing, a fundamental activity for writing proficiency. To improve the deaf's writing it would therefore be even more useful to propose exercises comparing the written forms of SL and of VL. However, this is not simple, as SLs are historically pure "oral" languages, without an established writing system. That notwithstanding, experiments using Sutton's (1995) SignWriting, a graphic representation system of SL, have shown how the knowledge of a SL writing system allows, in very natural ways, the emergence of metalinguistic reflections which can then be reinvested to better understand the structure and functioning of the own reference VL.

1. Introduction

Sign Languages (SL) are visual-gestural languages used by deaf[1] and hearing people who recognize themselves as members of the deaf community of their respective countries: in France, LSF is used; in the USA,

Claudia S. Bianchini   ⓘ 0000-0002-4783-1202

UFR Lettres & Langues, Bâtiment A3, 1 rue Raymond Cantel, TSA 11102, 86073 Poitiers Cedex 9, France

claudia.savina.bianchini@univ-poitiers.fr

1. The use of the word "deaf," instead of the more politically-correct "hard of hearing," is a deliberate choice of the author, which reflects the willingness of the signing deaf to be recognized as member of a linguistic community and not as people affected by a pathology.

ASL; in Italy, LIS; in Germany, DGS; in England, BSL (which has nothing to do with ASL although English is also spoken in the USA!); and so on[2], for a total of about 140 SLs worldwide[3]. SLs are not simple gestural transpositions of local vocal languages (VL), but they are real languages with their own syntax and lexicon, which allow to talk about any topic.

SLs have several characteristics that distinguish them from VLs, all due to the visual-gestural channel, which is the preferred one for the deaf (for further information: Cuxac, 2000; Cuxac and Antinoro Pizzuto, 2010; Sallandre, 2014). First of all, the meaning in SL is transmitted by the simultaneous use of multiple articulators (hands, arms, torso, head, mouth, eyes, etc.), all equally important, thus making the SL *multilinear* languages (in opposition to the *monolinearity* of the VLs, which have the mouth as sole articulator). These articulators move in the 4 dimensions of space-time, allowing SL to exploit all of them to create and syntactically organize the meanings: they are therefore *spatial* languages. This spatiality favors the use of *iconicity* for the creation of meaning, that is, the description of a selection of visual characteristics to describe entities or actions[4]. Iconicity influences not only the lexicon but the whole SL syntactic structure.

Although the whole SL lexicon has an iconic component, the value to be attributed to iconicity varies according to the type of syntactic structure used. Cuxac and Antinoro Pizzuto (2010) distinguish two kinds of structures, the Lexematic Units (LU) and the Transfer Units (TU), which differ from the signer's intention to just say (which in VL would be equivalent to saying "yesterday I ate a lot of pizza") or to say something while in the meantime showing it (which in VL would be equivalent to saying "yesterday I ate a piece of pizza this big" accompanying this sentence with a gesture to show the size of the pizza slice). In "saying without showing" (a.k.a. "visée non-illustrative"), the signer looks the interlocutor straight in the eyes and, using almost exclusively the hands, communicates the information using lexicalized units whose canonical form could be found in a SL dictionary. In "saying and showing" (a.k.a. "visée illustrative"), on the other hand, the signer looks at the space where the signs develop and, in so doing, activates both the

---

2. While these examples seem to imply that the mapping of countries into languages is one-to-one, this is not always the case: a country can have several SLs or one SL can be used in several countries, e.g., Swiss deaf utilize LIS (also used in Italy), LSF (also used in France) or DSGS (used only in German-speaking cantons).

3. https://www.ethnologue.com/subgroups/sign-language

4. However, it is necessary to note that iconicity does not constitute an obstacle to the ability of SLs to express abstract concepts; in these cases, in fact, a visual metaphor corresponding to the abstract concept will be rendered iconically: for example, to indicate the soul, the sign describes a subtle and light entity which goes towards the sky or, to indicate learning, the sign describes the act of taking entities and letting them enter in the head.

space and the signs themselves; furthermore, investing emself with the whole body (facial expression, head and torso movements, etc.) allows the interlocutor to see how a situation unfolded (that is called a "situation transfer"), how e acted and what a person looked like ("personal transfer"), or what physical characteristics typify an object or an entity ("transfer of size and shape"). The signs made with illustrative intention have long been considered pantomimic (Kendon, 1980), and for this reason these signs have struggled to be accepted as fully linguistic structures. However, Cuxac (2000) has demonstrated that these signs are linguistic elements with a structure and economy, and Sallandre (2003) has shown how, in some narrative forms, they can make up about 90% of what is signed.

Although the SLs origin is often associated with the moment of their institutionalization (i.e., the beginning of their use in schools for the deaf[5]), there are traces of their existence already in the writings of the philosophers of Greek antiquity, like Plato or Aristotle. Thus, SLs are in fact historical-natural languages with peculiar characteristics, used by specific communities, and which allow to express any concept which can be formulated in any other language. However, the gestural nature of these languages has often led institutions to ostracize them: in 1880, for example, the concluding declarations of the Congress of Milan led, throughout Europe, to a ban on SL use to educate deaf children; this prohibition lasted almost 100 years, devaluating SLs both in the hearing institutions and in the deaf community itself (Encrevé, 2012). Only starting from the '60s in the USA and then the '80s in Europe, with the movements of "deaf awakening," SLs were recognized as languages in all respects, starting a path of revaluation that continues (sometimes with difficulty) even today.

2.   Deafness, Literacy and Sign Language

Premise: in this section reference is made to France, where SL enjoys a relatively advanced linguistic recognition, especially since the promulgation of the Accessibility Law of 2005 which recognized the LSF as a language in all respects and which confirmed the right of deaf children's parents to choose how to educate their children. In Europe there are countries with even more advanced inclusive legislation, for example Austria, Finland and Hungary, which offer their SLs recognition in the

---

5. In France, for example, institutionalization took place between 1760 (the year of the foundation by the Abbé de l'Épée of the first school for the deaf in Paris) and 1791 (the year of the transformation of this school into the "Institut National des Jeunes Sourds" [INJS]); this school, still active, is known today as Institut Saint-Jacques or INJS-Paris.

Constitution; unfortunately, however, there are also countries with se-
rious delays, such as Italy, where the simple recognition of the linguistic
status of the LIS still seems far away and where the right to a bilingual
education is denied (although not forbidden, little or nothing is done
to favor it). Although countries offer their deaf children different edu-
cational possibilities, the following considerations can be applied to all
European countries (and beyond).

Today, profound deafness affects about 200,000 people in France and
LSF is practiced (more or less well) by about 80,000 deaf persons (not all
profoundly deaf) (Gillot, 1998). In most cases, deaf children are born into
hearing families, where they are the first and only deaf (Cuxac and Anti-
noro Pizzuto, 2010); nevertheless, the child's parents will have to choose
from birth which type of communication to favor. Analyses of children's
language development show that if a deaf child is exposed to SL early (ei-
ther because parents are deaf or because they decide to learn SL), eir un-
derstanding and production (in SL, not in VL!) at age two are comparable
to those of a hearing child of the same age (in VL, not in SL!) (Rinaldi et
al., 2014); if, on the other hand, oral communication is chosen, the child
will be able to discriminate the form of different words, but will show a
delay in both the development of comprehension and of oral production
(De Santis, 2010), due to the difficulty of learning an oral language with-
out being able to hear it. However, parents take the oral route more fre-
quently, since at the time of diagnosis they know little or nothing about
SL, and teaching the child to lip-read[6] and talk seems, alas, the simplest
way to promote eir social integration and educational success.

Whatever the language chosen in the family, right from enrollment in
kindergarten, French law (Assemblée Nationale and Sénat, 2005) allows
parents to choose for their children an education with or without LSF[7]:
the first case is called bilingual education, in where the "oral" language is
the LSF and the "written" language is French, the use of spoken French
being possible but not mandatory; the second case is called oral educa-

---

6. Lip-reading requires long training and continuous practice: it is a question of
discriminating what is said from the form of the mouth, but not all sounds results from
different mouth shapes. The deaf must not only be at a short distance and perfectly
in front of the speaker, but must also exploit clues from the facial expression, gestures
and general context. This activity is therefore considered by the deaf as very tiring
(DavSign, 2018).

7. This subdivision is here simplified to the extreme: there are many variations of
the "bilingual method," e.g., methods which more or less integrate oral French, which
have recourse to professors who teach in LSF or which use LSF interpreters who trans-
late what the professor says in French; in the same way, the "oral method" encompasses
many different educational systems, some of which prohibit any use of gestures while
others integrate gestural forms (in different phases and with different values) and even
involve sometimes the use of "true" signs (for a more complete view, refer to the thesis
of Leroy [2010]).

tion, in which French is used both for the written and the oral part (thru lip reading). However, this freedom of choice clashes with the scarcity of schools and institutions that guarantee the bilingual education. ANPES estimates that parents have chosen a bilingual education in 3,570 cases (out of 10,600 deaf children in school age), but that only 10% actually have access to a true bilingual class; for the remaining 90%, the way in which LSF is integrated into the school is not known (it could be just a purely oral education, associated with some course of LSF initiation).

Whatever the type of education chosen, the deaf child grows up in a world where VL is the primary environmental language and its written (and often oral) form is taught throughout school. Still, the assumptions that a deaf child has the same literacy skills of a hearing child of the same age, and that the number of illiterate adults among the deaf is in line with that of the hearing people are both incorrect. In fact, the rapport Gillot (1998)[8] indicates that 80% of deaf adults are unlettered, i.e., they have learned to read and write (they should be then literate) but do not have the necessary skills to understand and/or produce not even short texts. Most deaf people declare that they live negatively any situation in which writing is necessary, developing strategies aimed at avoiding recourse to writing and reading (Garcia et al., 2007).

How to explain this data? Set aside the theories of the early '900 concerning a deaf's phantomatic cognitive defect (for more details see Perini, 2013), the responsibility was then attributed to the use of LSF: however, since it was prohibited until the 1980s and is used by less than 40% of the deaf, it is unlikely that it could be responsible for such a debacle. Studying other possible explanations, Perini (ibid.) showed that the development of reading-writing skills is not so much linked to learning *how* to read and write (deciphering of letters, acquisition of phonological awareness) but to understanding *what* is written and read (knowledge of the world) and *why* (functions of writing). This link is also highlighted for hearing children, and in fact the discovery of both the world and the functions of reading and writing are part of the normal school curriculum since kindergarten ("Ministère de l'Éducation Nationale" 2015). For Perini, the development of reading and writing skills depends on the early constitution of a "favorable linguistic environment," in which the child's understanding is stimulated, rather than his linguistic production: it is therefore difficult to attribute the low level of literacy to LSF, especially since the number of deaf signers who claim to be at ease

---

8. It has been 30 years since this report, but deafness experts continue to cite its data because: 1) no similar study has been carried since then; 2) the situation, albeit improved thanks to the Accessibility Law of 2005 (Assemblée Nationale and Sénat, 2005), has not radically changed and, even if the number of deaf illiterates may have dropped, it is still about 10 times higher than that of the hearing people (a level estimated at 7% of illiterates in France in 2011 (ANLCI).

in writing is greater than that of deaf oralists (Garcia et al., 2007) and since in profound deaf children the knowledge of a SL smoothes acquiring writing skills (Niederberger, 2005).

With the same understanding of the world, however, the level of literacy of the deaf remains lower than that of their hearing peers (Dubuisson and Daigle, 1998). This statement is supported by the low number of profoundly deaf people who access the University, only 300 in 2016 (Micouleau, 2018). Another explanation can therefore be in the natural under-exposure of the deaf to the dominant VL, i.e., the language normally used to read and write: whether they have been educated with SL or without, the deaf cannot take advantage of the linguistic bath that every hearing child is exposed to (parents, relatives, television, passersby, all speak in front of them even when they are not talking *to* them). Although SL allows deaf signers to access content, the lower linguistic exposure to VL makes them less able to express themselves in VL; and, for the deaf oralists, the greater exposure to VL does not compensate for the problems related to the difficulty of lip-reading, a tough and tiring exercise that "occupies" the child in an effort of deciphering rather than understanding and which works only in mutual direct communication. Considering this lower exposure, some people (e.g., the researchers of the LSQ group at UQÀM[9]) have proposed to compare the deaf's difficulties in reading-writing not so much with that of their hearing peers, but with that of foreigners who are learning to read and write French.

3.   Role of Comparison Between Different Languages in Metalinguistic Development

In the '90s, Dabène and Ingelmann (1996) conducted an experiment involving two classes (about 50 children age 9–11 years) of an elementary school in Grenoble (France). These classes were characterized by a large presence of bilingual foreign children (speaking, in addition to French, Italian, Turkish, Arabic or Spanish); moreover, in these classes was active the "awareness of language" (Hawkins, 1987), a teaching methodology of Anglo-Saxon inspiration which, in the description of the authors, corresponds to:

> It is a question of arousing, in the child, from observations and manipulations carried out on a supporting language (L1)—French or a language of origin, but even on a set of languages as varied as possible—the awareness of what is the language universe in its variety, its functioning and its acquisition. We hypothesize that this type of work is likely to promote both the

---

9. Université du Québec à Montréal, Groupe de Recherche sur la LSQ et le Bilinguisme Sourd, `https://lsq.uqam.ca/`.

reasoned mastery of L1 and the learning of foreign languages, while integrating the contribution of the original languages of the alloglot children, which are thus legitimized." (Dabène and Ingelmann, 1996, p. 3, translated from French by the author)

For example, working on tenses, children were asked what they thought of a phrase like "tomorrow I disguise myself" and, letting them reflect in group, they were led not only to give the correct version "tomorrow I will disguise myself" but also to find how to justify to comrades why this form was correct, using examples based both on French and their native language. In another exercise, bilingual children were asked to explain to other children the difference between the morphology of the verbal system of their native language and that of French (of course, without using the word morphology, too complex for children of that age).

The authors stress the advantages of such an approach not only for bilingual but for monolingual children too: in fact, parallel to the improvement of French, bilingual children developed a reflection on the language of origin which, not being the language of schooling, was used without explicitly knowing its rules, while monolingual children discovered new linguistic dimensions hitherto unknown; furthermore, the use of their language in the classroom allowed allophone students to perceive their linguistic diversity as a richness for the whole class and not as an obstacle to their integration, effectively enhancing their language of origin.

Numerous studies on bilingualism (for a critical synthesis, see Besse, Marec-Breton, and Demont, 2010) have shown how knowing more than one language allows children to develop a strong metalinguistic awareness, that is "a subdomain of metacognition which concerns language and its use, in other words comprising: (1) reflections on language and its use; (2) the abilities to control and plan own linguistic processes" (Gombert, 1990, p. 27; translated from French by the author); it should be emphasized that developing a metalinguistic awareness does not necessarily mean knowing the specialized vocabulary necessary to express it (Dabène and Ingelmann, 1996).

In fact, deaf signer children are bilingual children, so why not try to use their knowledge of SL to improve their relationship with VL?

4.   Metalinguistic Awareness and Writing

The development of metalinguistic awareness is strongly linked with writing: apart from perhaps that made by Pānini[10], there is no systematic study of the organization of a language that has been done without

---

10. The Indian Pānini was the author of the first grammar of Classical Sanskrit, consisting of about 4,000 rules. The dating of his work is very controversial (Filliozat, 2020): on one hand, it describes the classical Sanskrit, which places it around the 7th

a system for writing that language; therefore, the linguist approaching a new language wonders, right from the start, about how to represent and transcribe it, in order to study it. According to Goody (1977), one of the major innovations resulting by the invention of writing is precisely that of allowing to reflect on language using the language itself.

Now, as illustrated in § 2, bilingual education for deaf children means written French and LSF replacing spoken French: the written form of LSF is not taken into consideration at all, although it would seem a good idea to use this knowledge to explain to children why French works in a certain way, or to show how the functions of writing are the same for SL and VL.

The reason for the total exclusion of written SL from school teaching is simple: there is no writing for SL! In fact, like most of the world's languages, none of the 140 SLs currently registered[11] has ever developed a writing system accepted and adopted by the deaf community, so it is not possible to use any SL writing system within a bilingual education. Since SL institutionalization, however, there have been several attempts to represent SL, some aimed mainly at linguistic research and others more focused at use in education.

5.   Historical Attempts at SL Writing

Attempts to represent SL can be divided into three broad categories (Bonnal-Vergès, 2008a), of which only the last one will be the scope of this paragraph: the design of the sign shape (Fig. 1a); the description in words of the sign shape (Fig. 1b) or of the image to which the sign seems to refer (Fig. 1c); the representation, through specially designed characters, of the parameters that make up the signs, which can be arranged linearly (Fig. 1d) or in a two-dimensional space (Fig. 1e). Drawings were, for a long time, the only way to try to represent SL because, as well summarized by Br. Louis, monk and educator of the deaf,

---

century BC; but on the other hand, the extent of his work is such as to seem impossible to be done without having recourse to writing, which appeared in the region of India where he lived only in the 3rd century BC. Supporters of older dating hypothesize that Pāṇini was able to compose his grammar without recourse to writing because he used the memory of his many disciples to "record" the rules he identified.

11. Ethnologue[12], an inventory of world languages, states that about half of the currently spoken languages have never developed a form of writing; of the other half, it is not possible to determine how many actually have a writing system that is still normally used and how many remain just purely oral, despite the existence of a graphic form. Given that Ethnologue's statistics are based only on the languages currently spoken, it can be said that, since humans developed the language, almost all the languages that have existed on earth have been solely oral.

12. `https://www.ethnologue.com/enterprise-faq/how-many-languages-world-are-unwritten-0`

(b) "Two-handed symmetrical sign. Hands are flat, slightly inclined from the horizontal plane and touch sideways. First, they go down and away, then continue to go down while keeping unchanged the distance between them"

(c) "Hands describe the roof and the external walls of a house"

(a)

(d)

(e)
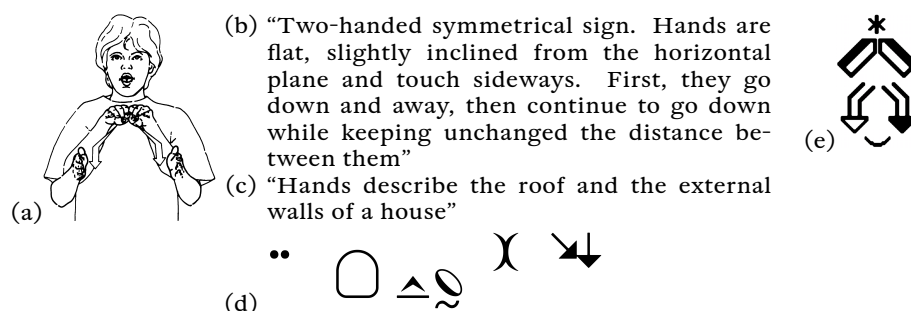
FIGURE 1. The sign [HOUSE] represented: (a) using a drawing (from Hanke, 2004); (b) describing its shape; (c) describing the picture to which the sign relates; (d) by a linear string of characters (HamNoSys); (e) by characters placed non-linearly (SignWriting)

> I would rather jump over the cathedral of Nantes mounted on a mule than Mr. Bouchet and company would create a universal signs dictionary without including drawings. (Bonnal-Vergès, 2008b, p. 140; translated from French by the author)

The first example of this type of representation is due to the French Auguste Bébian (1825), educator of deaf children, who had the intuition to describe SL through three series of symbols: one for the visible parts of the body; the second for the possible movements of these organs; the third for the facial expressions that accompany signs. In this way, the author hoped that "the deaf mute could express his thought on paper, as and more clearly than by gesture and without needing to translate it linearly into any language" (quoted by Renard, 2004; translated from French by the author).

This is therefore a first attempt to give deaf a way to write their language, without having to resort to written VL. However, this system was never adopted: at the time a movement against SL use in education was emerging in France, which led to Bébian's sacking and, 60 years later, to the prohibition of sign utilization which was sanctioned by the Congress of Milan in 1880.

Between 1880 and 1960, even in countries not directly affected by the Congress of Milan, such as the USA, SL underwent a process of devaluation, due to rapid medical-scientific progress which seemed ready to solve the problem of deafness at its root. However, in 1960, at Gallaudet University[13], linguist William Stokoe (Stokoe, 1960; Stokoe, Casterline,

---

13. Gallaudet University, founded in 1863, is the first and only University for the deaf in the world: the numerous courses offered, all in the field of Human Sciences and often in connection with deafness and SL, are provided exclusively in ASL. Contrary to Europe, the impact on the USA of the Congress of Milan was marginal, thus

and Croneberg, 1965) launched the revaluation process of ASL through the first modern study aimed at understanding the linguistic nature of SL (through the demonstration of their double articulation) and at creating an ASL dictionary. Stokoe considered that the minimum distinctive unit endowed with meaning (called *kinema* and not morpheme) corresponded to the whole sign, while the minimum distinctive unit without meaning (*cherema* and not phoneme) was to be identified in 4 "manual parameters": configuration (shape of the hand); orientation (direction of the carpus and metacarpus, this parameter was added only in 1965); place (area of the body in which the hand is located); and movement (action performed by the hand). He proposed to associate a character to each identified cherema and to organize his dictionary no longer on the alphabetical order of the English translation of the signs, but on the shape of the signs. To do this, it was necessary to be able to transcribe the signs with a common typewriter of the time: he therefore chose to use normal letters, numbers and mathematical symbols as characters, and to arrange them linearly according to a rigid syntax that would have allowed to put in "cheremic order" the signs in the dictionary. In this way, the Stokoe Notation (SN; Fig. 2a) was born, considered the first modern system for SL transcription.

SN could only encode ASL, since not only the names identified were those in use in the USA but the names chosen were in direct connection with the ASL: for this reason, numerous linguists (e.g., Bergman and Björkstrand, 2015; Kyle and Woll, 1985; Radutzky, 1992; Thoutenhoofd, 2003), proposed adaptations of SN but maintaining both the 4-parameter structure and the rigid linear formula; among these adaptations, HamNoSys stands out (Prillwitz, Leven, Zienert, and Hanke, 1989; Fig. 2b). Compared to SN, HamNoSys allows encoding a greater number of cheremas, through the use of basic symbols to which modifiers are added; it also allows representing some facial expressions; furthermore, the character design is iconically motivated, so that the system is exportable to other SLs and is easier to manipulate; finally, particular attention was paid to the computer integration of the system, so as to facilitate its use in the scientific field.

SN and its subsequent adaptations were tools designed by linguists for the sole purpose of research. On the contrary, SignFont (Newkirk, 1989; Fig. 3a), in addition to aspiring to be able to transcribe the SL, explicitly tries to establish itself as a writing system for SLs (Camurri and Volpe, 2003). However, the system has very few differences compared to HamNoSys: iconic characters, use of modifiers, the possibility of encoding some facial expressions, presentation in the form of a

---

making the existence of such a structure possible, despite the general process of SL devaluation. In the 1980s, a visit by a delegation of the French deaf to the Gallaudet University was among the triggers of the "ReveilSourd" movement in France.

$$\mathbf{B}_{\wedge >}\,\mathbf{B}_{\wedge <}\,\emptyset^{\div \wedge |}$$
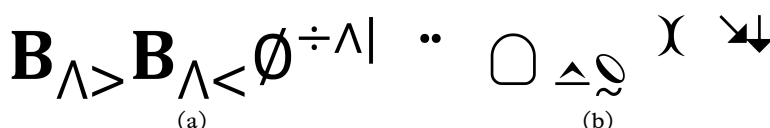
FIGURE 2.   The sign [HOUSE] transcribed with: (a) the Stokoe Notation;
(b) HamNoSys

$$\mathbf{\Pi \sigma \sigma}\,\times <> \,\sqrt{}$$
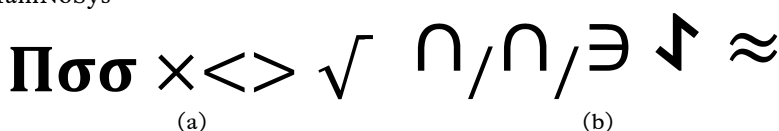
FIGURE 3. The sign [HOUSE] transcribed with: (a) SignFont; (b) the ASLphabet

rigid formula; the only difference is in the lower number of characters,
90 instead of HamNoSys's 200. This number was then further reduced
for ASLphabet (Supalla, McKee, and Cripps, 2014; Fig. 3b), a version of
SignFont developed for children to find signs in an ASL dictionary.

In addition to linguists, educators too tried to develop systems for SL
writing. Paul Jouison (1949–1991), an educator specialized in deafness,
was among the first (Jouison, 1995), to try to teach LSF to parents of
deaf children. During his first *cours de gestes* he decided to teach the LSF
lexicon and, separately, mime, which he thought was preparatory to the
acquisition of transfer units (TU, whose linguistic nature was still de-
nied). However, he soon realized that his method was not working and
decided to better understand the nature of TUs, trying to transform his
epilinguistic knowledge of LSF into metalinguistic proficiency useful
for transmitting LSF to other people. He therefore decided to film deaf
speakers in order to analyze both the SL lexicon and its syntax (ibid.):
the exercise not being possible without writing, he developed D'Sign
(Jouison, 1978), a transcription system that, unlike SN, was born with
the intent to transcribe the whole speech and not just isolated signs.
Like his predecessors, whose works he probably did not know (Jouison,
1995), he opted for a linear arrangement and did not choose particularly
iconic characters.

None of the systems presented so far have ever been used in schools
to permit deaf children to write their own language[14]. The reasons are
that they all have the same kind of problems:

---

14. However, the strong appreciation of some members of the French deaf commu-
nity for Jouison's precursor work, both with regard to D'Sign and the recognition of
SL as a true language, leads some educators of deaf children to hope for a reworking
of D'Sign, to be able to use it at school (e.g., SEB Poitiers: Service Education Bilingue
en Poitou-Charentes, `http://seb.poitiers.free.fr/seb.htm`).

m‹ ᶘp̣g̣v̀ùèʌ·Ⱡùⱳvz vè ℂp̣g̣v♦ùⱳvz·⌐ m‹
Iᶘq̣ɵ'g̣♦ɟxⱢ Iᴚᶘ́g̣vɵ̀q̣ús vè ℂp̣g̣v♦ùⱳvz·⌐ vé
φ̣̂'ᶘ́q̣u'ɛ̀ᴢ⌐ᴅ·ìvᶎᶈ vé ᶘ́ùuⱬ·ʌ ᴂ›ǡ'IgᶈⱢ ùᴂ›
ᶘ́ùuⱬv·φⱳⱳz' ᴂvé Iùuⱨvɵ́ᶈ·⌐v ɛᶎ̌ψɟu♦ɑ̣Iᶎ·
ʌⱢ·φⱳⱳz·ᴂvéɵ̀z·ψmɟ̀v   m‹   vé   φᴚ'ψxuᶘ́
·ĝᴂ›ʌⱢ·Iᶘψuɑ̣ɵmɟ̀v m‹ vé φ̣̂'ᶘ́ψᶎᶘⱳⱳⱳs·nφxɛψuⱢ
m‹ (φ̣̂'ᶘ́)ψᶎᶘⱳⱳⱳ·ψ̂'ᴂ›v φᶚ'ψuⱬ·qv Iᶘψuɑ̣ɵmɟ̀v
m‹ vé ⌴ᴂ̂x♦'ᶉʆⱳⱬ·vʌʌ zᶙᴂ̂x♦'ᶉʆⱳⱬ·vʌʌ·ψmɟ̀v
ǎ'ɛᶈʌⱳs   ɛᶎ̌ψɟu♦ɑ̣Iᶎ·ʌⱢ·φⱳⱳzɵ̀   ψmɟ̀èv
ℂp̣g̣v♦ùⱳvzè Iᶘψqeùⱳm̎ ℂᶉg̣g̣v·xmⱢvφ̀·è (mʌ
φ·ᶚ''ᶘu♢ⱬ·ɛⱼᶉɟq·ᶈ́v̂·ᶈᶋùs) vφ̀ φ̣·ᶚ''ᶘu♢ⱬ·ɛⱼᶉɟ
q·ᶈ́v̂·ᶈᶋùs vφ̀ φᶙʆⱼs ᴂñ›ǡ'ᶈᶎ ᴥ'ψxɑ̣ǎs φ̣·
ᶚ''ᶘu♢ⱬ·ɛⱼᶉɟq'φⱼ'ùsφⱳⱳz ᴂcφɵ'ψxɑ̣ǎsz·ñcψᴂsz
·ᴂcxǎᶎz φ·ᶚ''ᶘu♢ⱬ· ·ɛⱼᶉɟq·è ℂp̣g̣v♦ùⱳvz
ᴂ‹ɵ́·ᶘ́ψⱬv·ɵuu·ᴂɵɵ̀' ᴂᶎ·ᶘ́ψuuɑ̣v·ɵsɵxⱶ̄ Iᶘ́gvɵ'ℂᶉɟg̣
·mɑ̣ʌⱢv·ᴂ›v̄ʎ̃

A character walks in a swimming pool facility, buys a ticket which he exchanges for a coat hanger after thanking the receptionist. He takes off his clothes (sweater, pants, shoes), puts them on the hanger and brings it back to its place. He goes into the shower, after having checked its temperature with his foot. Coming out of the shower he goes through a gate and discovers a huge pool, along which people are running. He sees a diving board. A diver jumps and enters the water.

FIGURE 4. Example of a discourse in LSF transcribed using D'Sign; on the right, a non-literal translation of it

- Almost all of them have been developed to produce SL dictionaries, and in times when TUs were considered extra-linguistic idiosyncratic manifestations and the role of non-manual components (gaze, body movements, etc.) was considered marginal. Therefore, they offer a linear vision of SL (which, on the contrary, is multilinear) in which the facial expressions are considered marginal and the only units represented are the LUs, making impossible to code the signed speech. The only exception is D'Sign, but probably just because its inventor had no knowledge of the then linguistic theories.
- The choice of linearly arranging the different elements allows establishing an ordered set but in fact decomposes the time dimension (i.e., reading one character after the other) of an element that is perceived globally (since all parameters concur simultaneously to the sign realization), thus imposing an effort to decipher and recompose information.
- The choice of iconic characters should make it easier to read and learn a system, but these characters are often difficult to trace, making the writing process laborious; furthermore, the characters are not always well discriminated from each other, therefore they risk of being confused both in writing and in reading.
- The small number of characters, despite being an advantage in the learning phase, does not allow to distinguish between signs very similar but with different meanings. Reading therefore becomes not only a process of decrypting symbols but a work of re-association of meanings based on the context (when available).

All these linear systems lack the ability to provide not only a written but also an "intuitive" form of SL that allows children to focus on linguistic reflection and not just on deciphering the writing system. For

this reason it is necessary to develop a writable but, above all, easily readable system, able to represent the different characteristics of the language (multi-linearity, importance of both manual and non-manual components), this being the aim of another SL writing system, Sign-Writing.

6. SignWriting

SignWriting (Fig. 5) has been invented in 1974 by choreographer Valerie Sutton, already author of DanceWriting, a dance notation system. Contrary to the SL writing systems presented so far (D'Sign excluded), SignWriting was created with the intention of writing SLs and not of transcribing them. For this reason, great attention was paid to the reading process which is made simple by the use of the analogy between the signing and the writing spaces.

FIGURE 5. The sign [HOUSE] written using SignWriting

In SignWriting, (4-dimensional) signs are represented in a 2-dimensional vignette (x, y) where the width (x) and the height (y) correspond to those of the signer, while the depth (z) and the temporal deployment (t) of the sign are represented by graphic expedients. For example, a curved movement away, ascending and then descending, will be represented by ⤶ (see it larger in Fig. 6a): the use of the arrow allows to add information on the sign temporality, transforming in a simple drawing element the trace that the hand seems to design in space; the use of a thickening at the origin of the arrow (its "tail") allows to give indications about the depth of the sign, transforming it into mere perspective (the closer, the bigger). Fig. 6 shows this movement and its opposite, that is, a curved approaching movement, with the thickening towards the tip of the arrow (its "neck").

The vignette is therefore an analogical reduction of the signing space, like that found in drawings that were, and still are, used to create SL dictionaries. It takes into account all the parameters of SL, whether manual (configuration, orientation, position and movement) or non-manual (facial expressions, positions and movements of the body, shoulders and
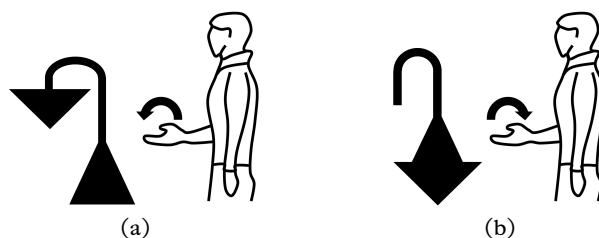
FIGURE 6. Upward and then downward curved movement that (a) moves away
or (b) approaches the signer's body

head, etc.), without a priori view on which components should or should
not be considered as participating in meaning transmission.

Still, SignWriting is not a simple drawing. Inside the vignette there
are characters, called SignWriting Symbols (SWsym) or glyphs, which
rigorously encode the way in which each occurrence of each parameter
must be represented: the straight movement of above will not be en-
coded by any arrow, but by a specific SWsym, which will be different
from that used to describe a shorter or more curved movement, a move-
ment opposite or directed in another direction, performed with the right
or the left hand.

The high number of SWsyms (over 37,000) is compensated by the
systematic nature of their organization (Table 1), since there are 432
"prototypical" SWsyms to which they are associated in a relatively rigid
way (Bianchini, 2012) of rules that allow to decline them and to obtain
the total number of possible SWsyms[15]. For example, a prototype rep-
resenting a configuration of the hand will require 4 rules: hand (the
right hand and the left hand are drawn mirrored); plane (the union or
separation between the dash representing the fingers and the shape rep-
resenting the palm makes it possible to distinguish whether the palm
is resting on the horizontal or vertical plane); orientation (the angle of
the glyph allows you to know the orientation of the hand on the plane);
color (the color of the hand, white, black or two-tones, makes it possi-
ble to distinguish which part of the hand is visible to the signer). Thus,
knowing the 4 rules that apply to all configurations, and knowing how to
design the prototype of a specific configuration, 96 SWsym can already
be used; knowing how to design the prototypes of the 242 configura-

---

15. While these rules are clearly stated Sutton, SW development over the years re-
sulted in many exceptions (Bianchini and Borgia, 2012). In her dissertation, Bianchini
(2012) has suggested a reclassification of SW that guarantees, without changing its
basic principles, an application without exception of the rules established by Sutton.

tions will permit 23,232 SWsyms, that is more than half of the existing SWsyms.

TABLE 1. Characteristics driven by different SWsym representing the configuration ⬚ (hooked index)[16]

SWsym	options								
hand	⬚ (right) ⬚(left)	D	G	D	G	D	G	D	G
plane	⬚ (vertical) ⬚ (horizontal)	V	V	V	V	H	H	H	H
orientation	⬚ ◇ ⬚ ◇ ⬚ ◇ ⬚ ◇ (0° to 315°)	0°	45°	90°	135°	180°	225°	270°	315°
side	⬚ (palm) ⬚ (side) ◼ (back)	P	C	D	P	C	D	P	C

The 432 prototypes are easy to learn, as almost all (402 out of 432) have an iconic connection with their referent: the hand SWsym looks like a hand, the mouth SWsym looks like a mouth, the SWsym of a movement looks like the trace left by the movement (Fig. 7).

flat and spread hand

flat and spread hand with hooked fingers

smiling closed mouth

open mouth with angles oriented down

segmented movement

curved movement
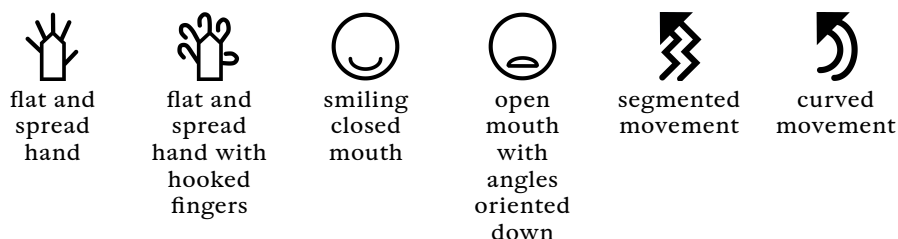
FIGURE 7. Examples of iconic SWsym prototypes

These characteristics of SW mean that 6−8 hours[17] are often enough for students[18] to acquire basic SW skills. Bianchini (ibid.) showed that

---

16. *Editor's note:* This character is called SIGNWRITING HAND-FIST INDEX BENT in the Unicode standard ("Sutton SignWriting Block," U+1D806).

17. 6 hours is the training time that a group of deaf signers required to produce the first text in SW (see infra); 8 hours is the time dedicated to SW classes within the degree course in "Sciences of language and French sign language," at whose end students are able to read isolated signs and write signs of simple movements.

18. The competence in SL greatly influences this figure: in the SW course taught by the author, students with a better SL level obtain better scores on both reading

the ease of SW use is linked both to its frequency of use, and to the subject's competence in SL.

Between 1998 and 2010 (Di Renzo et al., 2011), under the impulse of Elena Antinoro Pizzuto (EAP; 1951–2010), SW has been at the center of numerous linguistic studies, carried out at the ISTC-CNR in Via Nomentana in Rome (currently known as LaCaM), by a team called "Written-LIS Lab" ("Laboratorio di LIS Scritta"; LLISS) which brought together both deaf and hearing sign researchers[19]. Since the first days, SW has shown its ability to be learned very quickly, to be read easily and to stimulate metalinguistic reflection among the deaf who used it.

7. SignWriting as a Metalinguistic Tool

In 1998, after about 6 hours of SW tutorial, Tommaso Lucioli (TL) spontaneously wrote the short story "Home," which can be considered the first text in Written-LIS produced by a member of the LLISS. *Written-SL* (W-SL) is the formalization, directly in written form, of a thought in SL; it differs from *Face-to-Face-SL* (FF-SL), that is the formalization in "oral" SL of the aforementioned thought, which is then recorded and subsequently transcribed in order to analyze it. Within the LLISS, the system to formalize both W-SL and FF-SL is SW, but this could also be achieved with other SL writing systems: however, SW seems to allow, compared to other systems, a greater ease to develop metalinguistic reflections.

At the following meeting, TL proposed his text to the LLISS team then present (all deaf signers). The subsequent discussion led not only to the correction but also to the spontaneous analysis of different aspects of TL's text (Pennacchi, 2008): the combined use of different types of structures (i.e., [SNOW][20] is a LU that allows to say that there was snow,

---

and writing tests than colleagues with lower levels; moreover, the lack of SL knowledge seems to prevent even the simple decryption of the vignettes (i.e., their "reading aloud" without understanding the meaning).

19. SW was introduced to the ISTCCNR in 1998, when P. Rossini and B. Pennacchi, both deaf signers, self-taught SW using Sutton's manual (1995); later, they taught SW to 4 other deaf signers, T. Lucioli, A. Di Renzo, L. Ponzo and L. Lamano. The LLISS was created when the SW research received some funds in 2005. In 2007, SW was also taught to 3 graduate students, G. Gianfreda (deaf signer), G. Petitta and C.S. Bianchini (both hearing signers). All these people (but L. Ponzo), together with E. Antinoro Pizzuto, constituted the LLISS.

20. Conventionally, in the LS study, glosses are written in capital letters between square brackets. However, it should be emphasized that the use of glosses to describe SL is highly controversial (Pizzuto and Pietrandrea, 2001). The author of this article thinks that using a gloss in VL (a researcher's subjective interpretation, which "flattens" the SL on his/her own VL) without connecting it to the very sign appearance does not allow to take into account the peculiar SL characteristics and leads to pro-

while [THICK like that] is a TU that allows to show how the snow layer was); the role of the facial expression in differentiating signs with similar meaning (e.g., [FORCED] is associated with a particular shape of the mouth, the use of which makes it possible to identify this sign as "forced by the situation" and not "forced by a person"); the use of punctuation to define a set of signs as text (punctuation is typical of the written mode, in FF-SL no one would ever end a narration with [FULL STOP]).

	1	2	3	4	5	6	7
original text							
amended text							
verbal labels	HOUSE	MINE	SNOW	THICK (like this)	I	EXIT	IMPOS-SIBLE
translation	At home there is a very thick snow blanket: I can't go out						

	8	9	10	11	12	13	14
original text							
amended text							
verbal labels	FORCED	HOUSE	STAY	WORK	SIGN	WRITE	FULL STOP
translation	I have to stay at home: I'm working on written signs [Full stop]						

FIGURE 8. "House," the first narrative in Written-SL produced at the LLISS (author: T. Lucioli), in the original (no background) and corrected (grey background) versions

---

found biases in the analysis. The considerations carried out in this paper are therefore always made on the sign shape and not on its gloss. However, since it is unlikely that the reader of this article knows SW and LIS, for his/her convenience a *verbal label* (a term more appropriate than "gloss," according to Antinoro Pizzuto, 2008) is associated with the configuration, but always coupled with an image showing the sign form, and the label is used here for the sole purpose of facilitating understanding by the reader. Without adequate consideration for the sign structure, a verbal label is nothing more than the reflection of the (over)simplification/reduction into VL of a SL sign.

In the following months and years, countless texts of different lengths and types were written at the LLISS: real or fictitious stories; texts produced spontaneously or for needs related to LLISS research; texts in FF-SL or in W-SL. In particular, in 2009 the "Pear stories in LIS" corpus was created, consisting of different versions—both in FF-SL and in W-SL—of Chafe's "Pear Story" (1975). Whatever the nature of the texts produced, their sharing with the other members of the group has always led to a comparison of the older texts with the new ones and to spontaneously formulate reflections which gained in complexity and depth, as the group became more familiar with SW and competent in concepts of SL linguistics. The observation of this phenomenon led C.S. Bianchini, who reached the LLISS in 2007, to retrace the reflections made before his arrival (thanks to the consultation of the lab notebooks) and to stimulate new reflections through the analysis of texts produced within the LLISS (especially the "Pear" corpus): all these reflections are collected and examined in her doctoral dissertation (Bianchini, 2012) and will be summarized below. To these, new reflections will be added, reflections that emerged from the work of SL students (most hearing) who have followed the SW lessons of C.S. Bianchini from 2010 to today at the Université de Poitiers (France) and other training centers.

7.1.   Reflections on the Text Typographic Structure

The distinction between a shopping list, a Shakespeare's play, or a newspaper article is not only in its content but also in its form. Thus, part of the reflections that emerged at the LLISS focused on the question "what form must a text have to be considered a written text?". Already in "Home" we see that TL inserts [FULL STOP] at the end of his story, arguing that since it is a written text, it must necessarily end with a period. The choice is linked to the lack of knowledge, at the time, of the punctuation provided by SW, which provides SWsym for the full stop, comma, semicolon, and colon, exactly as in Latin writing. A few months later, after the team learned how to use these SWsyms, numerous discussions developed about the need for a comma in W-SL, or the use of colons. Furthermore, such discussions were related to the FF-SL too, leading the LLISS members to question whether it is right or not to attribute a punctuation mark to the different pauses present in the marked speech, but also to situations in which the pause is absent but punctuation could provide a complement of information (e.g., at the beginning of an enumeration, which would be announced in written VL by a colon and divided by semicolons).

Another recurring reflection concerns the management of the sheet space: for example, after discussing the need or not to give a title to narrative texts, numerous reflections arisen on how to highlight it with

respect to the rest. In fact, SW does not provide the possibility of under-lining, bold or italics. For this reason, LLISS members discussed different solutions on the frontier between orality (write "the title is XXXX") and writing (write the title horizontally instead of vertically like the rest of the text[21]; frame it; write it in another color; etc.).

7.2. Reflections on the Differences Between Writing and Transcribing

Since the beginning, the LLISS members note that SL writing implies considering the absence of immediate context and of the interlocutor but also the presence of co-text: the W-SL cannot therefore be a simple transposition of the FF-SL. This leads the team to speculate about "what differentiates FF-SL from W-SL?", so much so that a member asked EAP "are the books you read written or transcribed?".

Comparing different texts in W-SL and FF-SL, it initially seemed necessary to prefer LUs over TUs, since understanding of the former is less linked to the context. However, this choice would have distorted the SL, since in narrative texts TUs constitute about 90% of the signs produced (Sallandre, 2003)[22]. By comparing, in the "Pear" corpus, the same part of history but expressed in FF-SL and in W-SL it was possible to deepen the reflection, leading to the conclusion that it is not a question of replacing TUs with LUs, but of making more explicit the concepts expressed by TUs by using both LUs and TUs. For example, in the "Pear Story," a farmer collects pears by putting them in the pocket of his apron (Fig. 9). In FF-SL, this scene is described by the LU [PEAR] followed by a TU [PICK like so] and a unique TU that condenses [PUT like so PEARS IN THE POCKET like so PLACED there] (Fig. 9a). In W-SL, on the other hand, the same sequence is described by the same author by 3 TUs [CORD BEHIND THE NECK like so] [TRIANGLE IN FRONT like so] [POCKET like so], then by 2 LUs [EQUAL] and [KAN-GAROO], the renewal of the TU [POCKET] but used this time to say without showing (as if it were a LU), followed by the LUs [INSIDE] and [PEAR] and finally by the same condensed sign used in FF-SL [PUT like so PEARS IN THE POCKET like so PLACED there][23] (Fig. 9b). The explication can be done by using LUs which clarify the various parts

---

21. Apart the very first texts, such as "Home," the SW texts produced at the LLISS are written vertically, since the LLISS members had noticed that this made it easier to maintain the spatial references necessary to SL; also this choice is the result of metalinguistic reflections on the use of the sheet space.

22. The deaf's fear that SL writing may "distort" the language is often reported in research on the impact of a writing system on the deaf community (Bianchini, 2012; Garcia et al., 2007).

23. It is possible to notice slight differences in the choice of SWsymss to encode this sign. In fact, there is still no orthographic standard in SW, and is therefore possible to

of the TUs, by more detailed LUs and by procedures such as the use of metaphors and comparisons with other visual realities (e.g., "a pocket like that of a kangaroo").

(a) FF-SL	(b) W-SL		
PEAR	CORD BEHIND THE NECK like so	EQUAL	INSIDE
PICK like so	TRIANGLE IN FRONT like so	KANGAROO	PEAR
PUT like so PEARS IN THE POCKET like so PLACED there	POCKET like so	POCKET	PUT like so PEARS IN THE POCKET like so PLACED there

FIGURE 9. The introduction of the apron pocket by TL is (a) implicit in Face-to-Face-SL and (b) explicit in Written-SL (with approximated verbal labels)

A further examination of these reflections has led to the identification of situations in which it is necessary to pay particular attention to being explicit, as in the expression of emotions (usually conveyed orally by the facial expression alone, while in W-SL it is often necessary to make them explicit with a LU) or in the management of spatial references for deixis and anaphora (in FF-SL, deictic elements can be barely hinted with a finger or a head or body movement, while in W-SL they must be underlined and sometimes made explicit by LU or TU—especially in situation transfers).

---

write the same sign in slightly different ways: however, the reading of this sign will be identical.

7.3.   Reflections on the Concepts of Spelling, Standard and Error

In Fig. 9, the complex sign [PUT like so PEARS IN THE POCKET like so PLACED there] is identical in both FF-SL and W-SL. However, while the text in W-SL was conceived and written by the same author, the text in FF-SL was conceived and signed always by the same author but was transcribed by another LLISS member. Between the two "writers" (the writer of the W-SL text and the transcriber of the FF-SL narrative) it is possible to notice a difference in the choice of SWsyms used to encode the sign. These differences, highlighted from a search in the "Pear" corpus for signs with equivalent meaning but written and/or transcribed by different LLISS members, allowed the team to reflect on the question "how should a certain sign be written in SW?".

Since SW is a relatively new system, with a small number of users, for which Sutton itself neither imposes nor proposes strict rules, it is not possible to speak of SW "spelling". However, while sharing the texts, the LLISS members were able to state with absolute certainty that a vignette needed correction, and this was already evident from "Home". From the first discussions on the subject, the criterion of "readability" emerged: if the reader can easily read the sign and if what e signs (the form) and what e understands (the meaning) correspond to the intention of the author, then the sign is written correctly. However, this criterion involves "testing" each text with numerous readers in order to be sure that the own spelling is correct.

The reflections carried out in this area did not, in the end, lead to the definition of any SW "good spelling," but allowed to underline numerous things to do or to avoid in order to produce readable texts. For example, it is necessary not to overload the vignette with information (e.g., it is better to divide the sign into two vignettes or to leave out details that are not relevant for understanding); you must try to be redundant in the fundamental information (e.g., to indicate that the whole body moves to the right, it is better to place the indication of movement at both head and shoulder height); once chosen how to code a certain information, keep consistent; etc.

7.4.   Reflections on the Concepts and the Metalinguistic Vocabulary

The reflections on how to express the concepts also led the LLISS members to deepen the concepts of LU and TU, questioning on "how to distinguish a TU from a LU in writing?," e.g., [POCKET like so] from [POCKET] (see Fig. 9). According to Cuxac (2000), author at the origin of the distinction between "saying without showing" (i.e., the LUs) and "saying while showing" (i.e., the TUs), the gaze is different in the two types of structure: in the LUs it is directed towards the interlocu-

tor while in the TUs it is directed towards the hand or in any case the signing space. But in writing the interlocutor is absent, which makes it difficult to decide whether a gaze represented by a SWSYM is directed to the surroundings or at an imaginary interlocutor. For this reason, after several reflections arising from text misunderstandings (establishing a "showing glance" where there is none can completely change the meaning), a solution was developed: inventing a new SWSYM, a "look at the interlocutor" (Fig. 10a) to be inserted in each sign of LU type.

This reflection led the group to better understand concepts such as LU and TU, but also to question "how to provide SL with a specific terminology for such concepts?": while in the VL it is possible to import a word from one language to another, the change of expressive modality between VL and SL requires giving a "visual identity" to the concepts, which requires understanding them in depth. Thus, while at the beginning of the LLISS concepts like LU and TU were simply expressed through the realization of the letters U.L. and U.T. in manual alphabet, the subsequent metalinguistic reflections led to the creation of true signs for different linguistic concepts: "lexematic unit," "transfer unit," "iconicity," "visée illustrative" and other terms related to Cuxac's theory, but also "SignWriting," "Written-SL," "Face-to-Face-SL". The creation of a new sign requires a strong understanding of the different implications of the different concepts, which can only arise from the metalinguistic reflections made about the concepts themselves. Seeking how to establish a sign allowed to deepen the concepts, having a sign to express them allowed to simplify the manipulation of concepts and to strengthen their understanding, favoring the possibility of further reflecting on them.

Similarly, a sign has been attributed to different SWSYMs, so as to be able to more easily talk about the way an element is written: this is the case of the SWSYM "gaze looking at the interlocutor" (Fig. 10a), which represents two "i" resting at eye level and that soon was also used as a sign for "lexematic unit" (Fig. 10b).

A particular phenomenon is worth noting: in 2010, Cuxac and Antinoro Pizzuto decided to rename the "standard sign" to LU. Several sessions followed in which the LLISS members discussed, starting with the texts and theoretical explanations of EAP, on what this name change implied. Eventually, the sign "standard sign" became obsolete and was replaced by "lexematic unit" (first in the form of initials "L.U." and then as a real sign).

8.   Conclusions

The development of metalinguistic skills allows to pass from simply using a language to also understanding why it works in a certain way. Let-
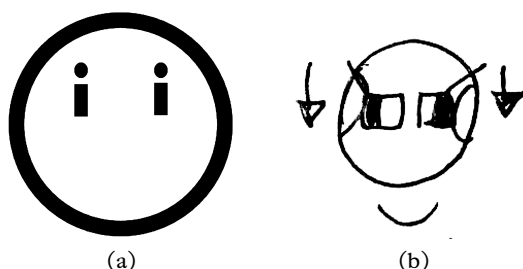
FIGURE 10. (a) tthe SWꜱʏᴍ invented by the LLISS members to write "gaze looking at the interlocutor" and (b) the sign associated first with this SWꜱʏᴍ and then with the concept of "lexematic unit" (in which the gaze is turned towards the interlocutor)

ting bilingual children to compare their languages has shown great benefits in acquiring these metalinguistic skills, and this is true whether you work on the oral or written part of the language. Numerous researchers have shown that deaf children are in a SL/VL bilingual situation, and that knowing an SL facilitates the VL knowledge too.

However, these analyses are always conducted in a trans-modal way, i.e., measuring how knowledge of the "oral" SL facilitates VL writing skills; of course, this is dictated by the absence of a written form for SL.

The various attempts to write SLs are inadequate to favor metalinguistic awareness as they omit fundamental parts of the sign (such as the non-manual components) and reduce a multilinear and threedimensional language to simple string, a one-dimensional representation. The only exception is SignWriting, which tries to propose an analog representation of the sign, showing both manual and non-manual components through a system that focuses on readability.

From its very first uses at the LLISS, it was possible to note that SW is a tool capable of stimulating metalinguistic activity both in those who write, and, above all, in those who read with it. Over the years, reflections have emerged on punctuation and notes, but also on layout and other formal characteristics that help define concepts such as sentences and paragraphs, together with the type of text under examination and the function of the different text parts. SW allowed also to reflect on the notions of orthography, standards, errors, even without the basic spelling rules of Western VLs. Its use has led LLISS members to reflect on both the lexical and syntactic structure of SL, highlighting the importance of non-manual components and the presence of nuances of meaning conveyed by small form variations. It also allowed to perceive the difference between writing and transcribing. Finally, SW brought out the need for a new terminology in order to formulate the metalinguistic reflections that it caused to spring out.

Thus, SW constitutes a valid tool to allow the deaf to compare the written modality of SL with that of VL, facilitating the understanding of the functions of writing, essential to successfully penetrate the realm of written VL.

However, the metalinguistic reflections collected by Bianchini (2012) in her thesis were almost all produced during the reading of SW. In fact, one of the "winning" features of SW, in terms of the development of metalinguistic skills, is its very high readability. Unfortunately, ease of reading does not go hand in hand with ease of writing, which turns out to be a long, laborious process and which, in the absence of clear spelling rules, requires continuous revisions to ensure correct readability of the text. In a school setting, therefore, the child could use SW to reflect on texts produced upstream by the teacher, but he could hardly use SW to write for himself; in the same way, the teacher could program the production of texts but could hardly improvise a text in SW during the lesson. SW is therefore a tool that can satisfy only a part of the functions that are required from a scholastic writing system. At present, this has led to exclude the use of SW in schools in almost all instances.

Nevertheless, in light of the advantages that the introduction of a form of SL writing may bring on the metalinguistic development of deaf children, one wonders whether it would not be worth trying anyhow to introduce SW in schools a.s.a.p., putting forward its potential above its limits, pending a more performing system (or a substantial evolution of SW to overcome the problem of its writability).

References

Antinoro Pizzuto, Elena (2008). "Meccanismi di Coesione Testuale e Strutture di Grande Iconicità nella Lingua dei Segni Italiana (LIS) e Altre Lingue dei Segni." In: *Atti Convegno Nazionale "La grammatica della Lingua Italiana dei Segni," Venezia, 16 Maggio 2007*. Venezia: Cafoscarina, pp. 16–17.

Assemblée Nationale and Sénat (2005). "Loi nº 2005–102 du 11 Février 2005 pour l'Égalité des Droits et des Chances, la Participation et la Citoyenneté des Personnes Handicapées." In: URL: http://www.legifrance.gouv.fr/affichTexte.do?JORFTEXT000000809647.

Bébian, Roch-Ambroise Auguste (1825). *Mimographie ou essai d'écriture mimique, propre à régulariser le langage des sourds muets*. Paris: L. Colas.

Bergman, Brita and Thomas Björkstrand (2015). *Teckentranskription: Report*. Stockholm: Institutionen för Lingvistik of Stockholms Universitet.

Besse, Anne-Sophie, Nathalie Marec-Breton, and Elisabeth Demont (2010). "Développement métalinguistique et apprentissage de la lecture chez les enfants bilingues." In: *Enfance* 10.2, pp. 167–199.

Bianchini, Claudia S. (2012). "Analyse métalinguistique de l'émergence d'un système d'écriture des langues des signes: SignWriting et son application à la langue des signes italienne (LIS)." PhD thesis. Université de Paris 8, Vincennes-Saint-Denis & Università degli Studi di Perugia.

Bianchini, Claudia S. and Fabrizio Borgia (2012). "Sign Languages: Analysis of the Evolution of the SignWriting System from 1995 to 2010 and Proposals for Future Developments." In: *Proceedings of the 3rd International Jubilee Congress of the Technical University of Varna*. Vol. 6, pp. 118–123.

Bonnal-Vergès, Françoise (2008a). *Sémiogenèse de la langue des signes française (LSF)*. Limoges: Éditions Lambert-Lucas.

——— (2008b). "Sémiogenèse de la langue des signes française: étude critique des signes de la langue des signes française attestés sur support papier depuis le XVIIIe siècle et nouvelles perspectives de dictionnaires." PhD thesis. Université de Toulouse II.

Camurri, Antonio and Gualtiero Volpe (2003). "The Development of a Computational Notation for Synthesis of Sign and Gesture." In: *Selected Papers from Gesture-Based Communication in Human-Computer Interaction, 5th International Gesture Workshop*. Ed. by Crombie K. Smith and W. Edmondson, pp. 312–323.

Chafe, Wallace L. (1975). *The Pear Story: the Movie*. Berkeley: University of California. URL: http://www.youtube.com/watch?v=bRNSTxTpG7U.

Cuxac, Christian (2000). *Langue des signes française (LSF): les voies de l'iconicité"*. Paris: Ophrys.

Cuxac, Christian and Elena Antinoro Pizzuto (2010). "Émergence, norme et Variation dans les langues des signes: vers une redéfinition notionnelle." In: *Sourds et langues des signes: norme et variations*. Ed. by B. Garcia and M. Derycke. Vol. 131. Langage et Société, pp. 37–53.

Dabène, Louise and Christèle Ingelmann (1996). "Un multilinguisme en construction: l'éveil de la conscience métalinguistique." In: *AILE Acquisition et Interaction en Langue Étrangère* 7, pp. 123–138.

DavSign (2018). "Lire sur les lèvres, c'est facile pour les sourds et malentendants." URL: http://www.supersignes.com/blog/lire-levres-sourds-malentendants/.

De Santis, Daniela (2010). "Lo Sviluppo del Linguaggio nel Bambino Sordo e Udente: Due Modalità Comunicative a Confronto." In: *Studi di Glottodidattica* 1, pp. 75–91.

Di Renzo, Alessio et al. (2011). *Scrivere la LIS con il SignWriting: Manuale Introduttivo*. Rapporto tecnico del Progetto FIRB-VISEL. Roma: CNR. DOI: doi:10.13140/RG.2.1.4079.6001.

Dubuisson, Colette and Daniel Daigle, eds. (1998). *Lecture, écriture et surdité—Visions actuelles et nouvelles perspectives*. Montréal: Les Éditions Logiques.

Encrevé, Florence (2012). *Les sourds dans la société française au XIXe siècle: idée de progrès et langue des signes*. Grâne (Drôme, France): Créaphis.

Filliozat, Pierre-Sylvain (2020). "Pāṇini (ve s. av. J.-C. env.)" In: *Encyclopædia Universalis*. URL: https://www.universalis.fr/encyclopedie/panini/ (visited on 28/09/2020).

Garcia, Brigitte et al. (2007). *Rapport du Projet RIAM-ANR LS Script, 2005–2007*. Agence Nationale de la Recherche.

Gillot, Dominique (1998). *Le droit des sourds: 115 propositions. Rapport au premier ministre*. La Documentation Française.

Gombert, Jean Émile (1990). *Le développement métalinguistique*. Paris: Presses Universitaires de France.

Goody, Jack (1977). *The Domestication of the Savage Mind*. Cambridge: Cambridge University Press.

Hanke, Thomas (2004). "HamNoSys: Representing Sign Language Data in Language Resources and Language Processing Contexts." In: *Workshop on the Representation and Processing of Sign Languages, 4th International Conference on Language Resources and Evaluation (LREC Lisbon)*, pp. 1–6.

Hawkins, Erik (1987). *Awareness of Language: an Introduction*. Cambridge: Cambridge University Press.

Jouison, Paul (1978). "Cours de Gestes 1977–1978." In: *Bulletin de l'Association Ferdinand Berthier*. reprinted: 2008. *Archives de Langue des Signes*. Limoges: Lambert Lucas, pp. 1–110.

———— (1995). *Écrits sur la Langue des Signes Française (LSF)*. édition critique établie par Brigitte Garcia. Paris: L'Harmattan.

Kendon, Adam (1980). "A Description of a Deaf-Mute Sign Language from the Enga Province of Papua New Guinea with Some Comparative Discussion. Part I: The Formational Properties of Enga Signs; Part II: The Semiotic Functioning of Enga Signs; Part III: Aspects of Utterance Construction." In: *Semiotica* 31.1/2 & 32.1/2–3/4.

Kyle, Jim G. and Bencie Woll (1985). *Sign Language: the Study of Deaf People and their Language*. Cambridge: Cambridge University Press.

Leroy, Élise (2010). "Didactique de la langue des signes française, langue 1, dans les structures d'éducation en langue des signes: attitudes et stratégies pédagogiques de l'enseignant sourd." PhD thesis. Université de Paris 8 Vincennes-Saint-Denis.

Micouleau, Brigitte (2018). "Question écrite: accessibilité des étudiants sourds à l'enseignement supérieur." URL: https://www.senat.fr/questions/base/2018/qSEQ180203125.html.

"Ministère de l'Éducation Nationale" (2015). In: *Éduscol: Ressources maternelle—Mobiliser le langage dans toutes ses dimensions. Partie III. 1: L'écrit. Découvrir la fonction de l'écrit*. URL: https://cache.media.eduscol.education.fr/file/Langage/40/0/Ress_c1_langage_ecrit_fonction_456400.pdf.

Newkirk, Don (1989). *SignFont Handbook*. Bellevue, WA: Edmark Corp.

Niederberger Nathalie et Prinz, Philip (2005). "La connaissance d'une langue des signes peut-elle faciliter l'apprentissage de l'écrit chez l'enfant sourd?" In: *Enfance* 57.4, pp. 285−297.

Pennacchi, Barbara (2008). "Mettere Nero su Bianco la LIS." In: *I Segni Parlano: Prospettive di Ricerca sulla Lingua dei Segni Italiana*. Ed. by C. Bagnara et al. Milano: Franco Angeli, pp. 140−147.

Perini, Marie (2013). "Que peuvent nous apprendre les productions écrites des sourds? Analyse de lectes écrits de personnes sourdes pour une contribution à la didactique du français écrit en formation d'adultes." PhD thesis. Université de Paris 8 Vincennes-Saint-Denis.

Pizzuto, Elena and Paola Pietrandrea (2001). "The Notation of Signed Texts: Open Questions and Indications for Further Research." In: *Sign transcription and database storage of Sign information*. Vol. 1/2. Sign Language Linguistics, pp. 29−43.

Prillwitz, Siegmund et al. (1989). *Hamburg Notation System for Sign Languages: an Introductory Guide*. Hamburg: Signum Press.

Radutzky, Elena (1992). *Dizionario Bilingue Elementare della Lingua Italiana dei Segni: oltre 2500 Significati*. Roma: Edizioni Kappa.

Renard, Marc (2004). *Écrire les signes: la mimographie d'Auguste Bébian et les notations contemporaines*. Les Essart-le-Roi: Éditions du Fox.

Rinaldi, Pasquale et al. (2014). "Sign Vocabulary in Deaf Toddlers Exposed to Sign Language since Birth." In: *Journal of Deaf Studies and Deaf Education* 19.3, pp. 303−318.

Sallandre, Marie-Anne (2003). "Unités du discours en langue des signes française: tentative de catégorisation dans le cadre d'une grammaire de l'iconicité." PhD thesis. Université de Paris 8 Vincennes-Saint-Denis.

——— (2014). "Compositionnalité des unités sémantiques en langues des signes: perspective typologique et développementale." Habilitation à Diriger des Recherches. Université de Paris 8 Vincennes-Saint-Denis.

Stokoe, William C. (1960). "Sign Language Structure: an Outline of the Visual Communication Systems of the American Deaf." In: *Journal of Deaf Studies and Deaf Education* 10, pp. 3−34.

Stokoe, William C., Dorothy Casterline, and Carl Croneberg (1965). *A Dictionary of American Sign Language on Linguistic Principles*. Reprinted: 1976. Burtonsville MD: Linstock Press. Washington, DC: Gallaudet College Press.

Supalla, Samuel, Cecile McKee, and Jody Cripps (2014). "An Overview on the ASL-phabet." In: *Gloss Institute's Monograph Series*. Vol. 1, pp. 1−18.

Sutton, Valerie (1995). *Lessons in SignWriting*. La Jolla, CA: Deaf Action Committee.

Thoutenhoofd, Ernst (2003). "The British Sign Language Variant of Stokoe Notation: Report on a Type-Design Project." In: *Sign Language Studies* 3.3, pp. 341−370.

Language Identity
Through Cyrillic Script

From Romanian to Moldovan by Automatic Transliteration in the Wikimoldia Project

Christian Koch

Abstract. Wikimoldia was a website active from 2018 to 2019 on which the Romanian Wikipedia was transliterated into Cyrillic by using a PHP script. This paper discusses the technical background of the automatic transliteration performed in Wikimoldia and links the project of a Cyrillic-language Romanian Wikipedia to the political and linguistic controversies surrounding the status of the Moldovan language. It discusses how Wikimoldia can benefit the Cyrillic-socialised minorities in the eastern periphery of Romanian-speaking areas. The use of machine transliteration can also be of interest in the context of other multialphabetic languages.

1. Introduction

Wikimoldia sounds like the name of another of Wikipedia's many offshoots (cf. Wikimedia, Wikisource, etc.), but alludes to a geographical area, namely the Republic of Moldova. If one tries to open the page `http://wikimoldia.org` today, there will not be anything of what was offered here between September 2018 and September 2019: a Romanian Wikipedia in Cyrillic script.

The aim of this article is to reconstruct the technical functioning of this automatically generated page and to discuss the potentials and difficulties of the project. Furthermore, the paper deals with the sociolinguistic background that may have motivated the Wikimoldia project. Understanding linguistic diversity in Wikipedia as a contribution to minority language vitalisation (cf. Born, 2007; Coulmas, 2018, 198f.), Wikimoldia can also raise the question of an attempt to vitalise this language, however it has to be defined at the same time what kind of language Moldovan actually is. From a linguistic point of view, the term 'Moldovan language' is highly problematic and requires a critical analysis, not least in view of the fact that, even within Romance linguistics,

---

Christian Koch   0000-0002-6697-3468

Romanistik / Angewandte Sprachwissenschaft und Didaktik, Universität Siegen, Adolf-Reichwein-Str. 2, AR-IF 229, D-57076 Siegen, Germany

E-mail: koch@romanistik.uni-siegen.de

the Romance-speaking varieties in the far east of Europe are among the rather unknown territories. The first part of the article is devoted to this aspect, before the following sections take a closer look at Wikimoldia from a technical and functional perspective.[1]

2.   On the Status of the Moldovan Language

2.1.   Stages of Language Naming

The term 'Moldovan language' (Rom. *limbă moldovenească*) has a different meaning in different political contexts and time periods. Gabinskij (2002, p. 133) indicates 'Moldovan' as a non-scientific everyday term[2] for the language of the Republic of Moldova, but also as a subglottonym of the glottonym 'Romanian' (ibid., p. 139), which may seem acceptable from a scientific point of view, if one speaks of a (geopolitical) variety of Romanian rather than of a language of its own. However, this also seems problematic because in dialectological descriptions the Moldovan dialect is understood as a geolectic area north of the Daco-Romanian dialect (cf. Olariu, 2017, p. 108) and this geolectic area is largely situated within the political borders of Romania.

The territory of the modern Moldovan Republic had an eventful history along the 20th century: Bessarabia, previously part of the Russian Empire, became in its majority a part of Greater Romania in 1918, then in 1944 it converted into the Moldovan Soviet Socialist Republic, before the independent Republic of Moldova was founded in 1991. The affiliation to the Russian Empire and the Soviet Union is still visible today in the presence of the Russian language and the Cyrillic alphabet. Since its foundation, the present state has been divided politically and, to a certain extent, linguistically into two regions on both sides of the Dniester River. Transnistria as a breakaway republic with half a million inhabitants in eastern Moldova confronts the country with a conflict which is still insoluble and hinders the integration into European institutions and in the rapprochement with Romania. The ongoing linguistic separation is based on the use of the Latin alphabet, as—with the exception of Transnistria—the Latin alphabet has been reintroduced in the independent Republic of Moldova: "decretarea limbii române ca limbă de stat și reintroducerea alfabetului latin, din 3 noiembrie 1990"[3] (cited in

---

1. This article a slightly extended and revised English translation of Koch (forthcoming).

2. German original: "(nichtwissenschaftliche) Alltagsbezeichnung".

3. Translation: the decree of Romanian as the national language and the reintroduction of the Latin alphabet, from 3 of November 1990.

Cimpoeșu and Musteață, 2018, p. 50). While the Declaration of Independence that is quoted here still refers to the Romanian language, a new decree was issued in 1994: "limba de stat a Republicii Moldova este limba moldovenească"[4] (cited in Olariu, 2017, p. 22). This can be understood as a return to Soviet identity construction which the Romanians Dorin Cimpoeșu and Sergiu Musteață accuse in harsh words:

> au fost legiferate tezele staliniste false despre apartenența etnică și lingvistică a populației românești prin introducerea în legea fundamentală a sintagmelor 'limbă moldovenească' și 'popor moldovenesc' contrare adevărului științific și istoric[5] (Cimpoeșu and Musteață, 2018, p. 61)

However, one can also more moderately assume a national identity building in which Moldova breaks away from its position as Romania's satellite state and promotes linguistic independence with its own glottonym. A bit later, the national anthem entitled *Limba Noastră* has also been established, which, along with the national holiday *Limba Noastră (cea Română)*, emphasises the outstanding importance of the national language in Moldova. The text of the anthem, which goes back to a much older poem by Alexei Mateevici (1888–1917), uses the politically more neutral possessive determiner ('our language') instead of a glottonym.

As an official language, Moldovan had a coding in the ISO 639 standard as "mo"/"mol," which was however already abolished in 2008 (cf. `https://iso639-3.sil.org/code/rum`). Finally, in 2013 it was decided to revert to the designation *limbă română* in official language use, probably also in order to strengthen the ties with Romania and thus the bridge to the European Union. This step was justified by the designation in the declaration of independence:

> prevederea conținută în Declarația de Independență referitoare la limba română ca limbă de stat a Republicii Moldova prevalează asupra prevederii referitoare la limba moldovenească conținute în articolul 13 al Constituției[6] (cited in ibid., p. 24).

In Transnistria, the renaming of the language was not applied, but Moldovan in Cyrillic script remained the official language (along with

---

4. Translation: the official language of the Republic of Moldova is the Moldovan language.

5. Translation: false Stalinist theses on the ethnic and linguistic affiliation of the Romanian people were established by law, introducing in the basic law the expressions "Moldovan language" and "Moldovan people," which contradict scientific and historical facts.

6. Translation: the provision on the Romanian language as the official language of the Republic of Moldova contained in the Declaration of Independence takes precedence over the provision on the Moldovan language in Article 13 of the Constitution.

Russian, Ukrainian and, regionally, Gagauzian). It should also be men-
tioned here that the Romanian language area extends beyond the east-
ern border of Moldova into Ukraine where the Romanian-speaking
minorities in northern Bucovina and the Ukrainian part of Transnis-
tria (or Bessarabia), but less frequently in Transcarpathia, feel close
to Moldovan identity (cf. Dahmen, 2018, 345ff). Due to the "of-
ficial" language names used today, the name *limbă moldovenească* (or
лимбэ молдовеняскэ) is increasingly narrowed down to the language of
Transnistria and the mentioned regions of Ukraine. The justification
for the use of this glottonym is linked to politically sensitive Transnis-
trian normative concepts. From a scientific point of view, the term
'Moldovan' can be used in this context as a functional term for 'Cyrillic
Romanian' without any political classification. The fact that it is pri-
marily the writing system and less the diatopic variety that is decisive
has to do with the way in which transliteration is carried out. This will
be the subject of the next section.

2.2.   Principles of the (New) Cyrillic Script of Romanian

The adjective *new* indicates that the principles of Cyrillic script in the
20th century are not related to the use of Cyrillic in early Romanian
writing since the 16th century, because Latin and Cyrillic letters had
already been in competition with each other for more than 300 years
before the Latin alphabet became established (cf. Onu, 1989). With the
beginning of Joseph Stalin's rule over the Soviet Union, the Cyrillic al-
phabet became the identity-giving symbol for the majority of the re-
gional languages in USSR. In the part of Transnistria not belonging to
Greater Romania, the Cyrillic script was introduced in 1928 for Roman-
ian writing and, with a brief interruption between 1933 and 1937, be-
came firmly established (cf. Kramer, 1989, p. 15). As mentioned above,
Moldova then became Soviet in 1944 and the Cyrillic alphabet was in-
troduced throughout the Soviet Republic, where it remained until inde-
pendence in 1991.

  The most important principle of the Cyrillic transliteration[7] from
1928 onwards was—as with all newly written languages of the Soviet
Union—the greatest possible harmony with the phoneme-grapheme cor-
respondences of Russian. However, the languages were also granted a

---

  7. According to Zikmund (1996, p. 1592), 'transliteration' is understood as the
language-indifferent transliteration opposed to target language-specific transcrip-
tion. In this distinction, the term 'transcription' would also be conceivable, since
numerous specific features of Russian are present in Cyrillic writing. However, un-
like the transcription of proper names, for example, it is not a form of integrating
written forms into the target language Russian, but rather the transcription of an en-
tire language system, which is why we prefer the term of transliteration.

certain degree of autonomy, so that their graphemes did not need to reflect all the special features of Russian. Individual letters of the Russian alphabet were converted and new letters were introduced—exactly one in the case of Romanian. This will be illustrated by a few examples (Tab. 1).

TABLE 1. Examples of Romanian letters in Cyrillic script

Example	Sound	Russian	„Moldovan"	Romanian
(1)	[ɨ]	ы	ы	â, î
(2)	[ʲ]	ь	ь	i
(3)	[ə]	(э $\hat{=}$ [ɛ])	э	ă
(4)	[d͡ʒ]	–	Ӂ	g (+e/i)
(5)	[iʲa]	ия	ия	ia
(6)	[l]	ль	л	l

1. The unrounded closed central vowel [ɨ] is characteristic of both Russian and Romanian, and the Cyrillic writing can solve the problem of the two graphemes <â> and <î> of Romanian.
2. The letter <i> is used in Romanian to indicate, inter alia, the palatalisation of consonants which are syllable-final or usually word-final. In some Slavic languages, the so-called soft sign <ь> follows on the palatalised consonants. The difficulty of transliteration due to the different functions of <i> is explained below.
3. For the Romanian language, the shwa sound written as <ă> is characteristic and not uncommon even in stressed syllables. Russian knows the sound only in unstressed syllables as a reduction level of /a/ and /o/. At this point, an imitation of Russian grapheme-phoneme correspondence would hardly be possible, and instead the third-last letter <э> of the Russian alphabet is used, whose Russian phonological value /ɛ/ is not necessary for the representation of a specific sound in Romanian.
4. Romanian and Gagauzian have got their own letter for the affricate /d͡ʒ/: <Ӂ>, which is distinguished by a diacritic breve from <ж>, because <ж> as the transliteration of Rom. <j> is also needed. This shows the preference in Cyrillic scripts for diacritics over digraphs, although a spelling like <дж> would be intuitively easier to read.
5. In various Slavic languages, iotation plays an important role, i.e., some vowels have an approximate initial [j]. Russian has its own letters for this purpose (<ю>—[ju], <я>—[ja], possibly also <ё>—[jo]). The letter <e>, which is identical in both alphabets, is regularly used with iotation in both Russian and Romanian. Another special feature of Russian is the graphic marking of the intervocalic iotation, which

is also conceivable as an epenthesis in Romanian, but is not graphically marked. As the last letter in the name of *Викимолдия* shows, the transliteration is based on the graphic representation of the epenthesis according to the Russian model.

6. However, the name *Викимолдия* shows that <л> can work without a soft sign for the articulation of the lateral [l], since the articulation of velarised (or hard) [ɫ] is irrelevant for Romanian.[8]

Remarkable about the transliteration rules developed during the Soviet era is that regional peculiarities of the articulation of the Moldovan variety were not taken into account. Instead, the pronunciation standard of Daco-Romanian (cf. Gabinskij, 2002, p. 135) is fully applicable and no attempt has been made—as it has happened with other Soviet regional languages—to represent a variety in Cyrillic that did not have any established writing system before. The demarcation from Romania rather occurred in the field of lexis as a rejection of the so-called *limbă păsărească* ('bird language'), which denotes the more sophisticated language oriented towards the Romanian norm (cf. ibid., p. 135).

3.   Key Data on Wikimoldia

The site `http://wikimoldia.org` was launched in September 2018 and was then accessible for one year. We can presume that the termination of the online presence happened due to the fact that the contract of use for the domain was not renewed. Therefore, direct access to the URL is no longer possible. Via the large web archive *Wayback Machine* (`https://web.archive.org/web/*/http://wikimoldia.org/*`), 520 pages are still available, but this is only a fraction of the several hundred thousand pages once available.

Wikimoldia can be seen as a phantom page to the Romanian Wikipedia, because the contents of Wikipedia are transliterated into Cyrillic with the help of a PHP script, that we are going to discuss in more detail in the following section. In addition, the name *Wikipedia* is replaced by *Викимолдия*, so that the artificiality of the pages is not obvious at first glance (Figs. 1 and 2).

In the top left-hand corner, it is noticeable that the signature of the logo "Wikipedia / Enciclopedia liberă" has not been transliterated or replaced, as this is a graphic element. The exact layout and hypertext structure are transferred to Wikimoldia; all hyperlinks to articles work and lead to corresponding pages in Wikimoldia. However, the search field at the top right is not functional (Fig. 3).

---

8. This does not mean that in Russian there would be a soft sign at this point. Rather, unlike in Romanian, *Молдавия* is articulated with velarised [ɫ].
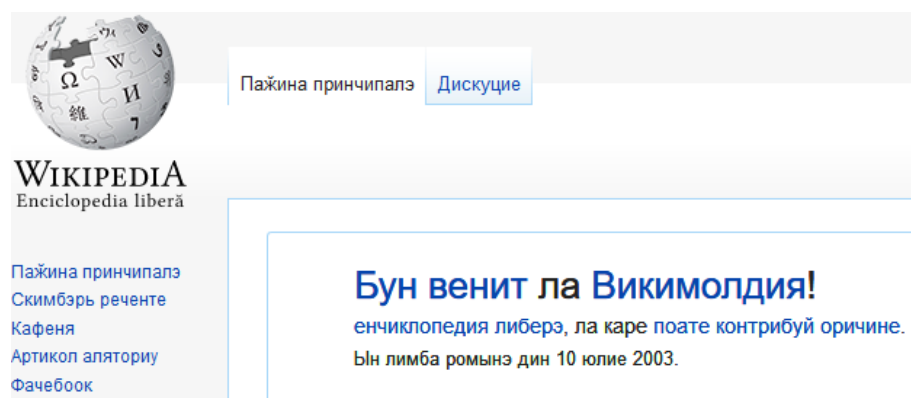
FIGURE 1. Partial screenshot of the Wikimoldia homepage
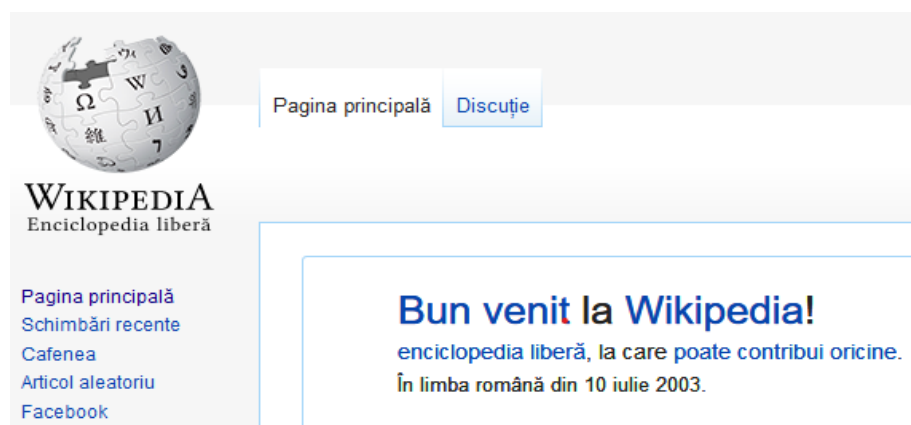
FIGURE 2. Partial screenshot of the Romanian Wikipedia homepage (ro. wikipedia.org)

It is only possible to enter words in Latin script and this will bring the user to the Romanian Wikipedia. Thus, in the sense of Leca-Tsiomis (2006), the encyclopaedic order works by the hypertextual reference structure, but not the alphabetical order in the form of direct look-up. In Wikimoldia it was only possible to call up an article in a targeted manner by manually replacing "ro.wikipedia" with "wikimoldia" in the URL of the Romanian Wikipedia article.

Before discussing the functionality and potentials of Wikimoldia in more detail in the following sections, the question should be allowed whether Wikimoldia could actually represent something like a vitalisation attempt, i.e., in particular whether the Cyrillic transliteration of
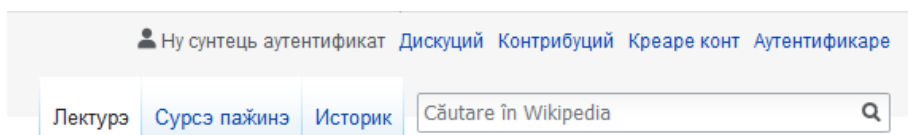
FIGURE 3. Search field "Căutare în Wikipedia" in Wikimoldia

the Romanian Wikipedia was seriously used. In June 2019, the usage data analysis of SimilarWeb (https://www.similarweb.com) still provided numbers for the period May to March 2019 (Tab. 2).

TABLE 2. Wikimoldia user numbers according to *SimilarWeb*

Total Visits (March-May 2019)	46,122
Monthly Visits	15,374
Monthly Unique Visitors	< 5,000
Avg. Visit Duration	00:01:07
Pages / Visit	1.53

In the three months, the website was accessed 46,122 times, with the number of users estimated at less than 5,000. On the basis of the described diversion via the Romanian Wikipedia articles, one could assume that the number of actual users could be higher, since—unlike in Wikipedia—it was not possible to access concrete articles via the Wikimoldia homepage. *SimilarWeb* also made it possible to determine from which countries users accessed the site (Tab. 3).

TABLE 3. User proportions by country according to *SimilarWeb*

	Country	Traffic Share	Country Rank
1	Ukraine	33.36%	·148,316
2	Russia	16.31%	·722,063
3	Moldova	13.02%	·42,216
4	Turkey	10.21%	·1,153,092
5	United States	4.39%	·1,441,191

According to this data, Moldova is only in third place after Ukraine and Russia. This may be simply because Ukrainian and Russian IP addresses are used in Transnistria. But it can also be linked to the Romanian-Moldovan minorities mentioned above, especially in Ukraine. Moreover, the higher number of users from these coun-

tries could also have something to do with the Cyrillic script, which builds a bridge to Romanian here. In August 2019, for example, `https://news.ru` quotes Wikimoldia as the source for a photo of the former Romanian Education Minister Ecaterina Andronescu (`https://news.ru/europe/slova-ob-iznasilovannoj-devochke-stoili-rumynskomu-ministru-dolzhnosti/`). It is probable that a Russian reporter in the search for a picture of "Екатерина Андронеску" has just found a relevant result via Wikimoldia. This example, as well as the user figures mentioned above, provide rough indications of a certain vitality of use for the period during which Wikimoldia was online.

4. Automatic Transliteration in Wikimoldia

4.1. Operating Modes

For the automatic transliteration of Romanian into Cyrillic letters according to the principles described above, the PHP script `slava37md2` was published in *GitHub* in August 2018 (`https://github.com/slava37md2/wikimoldia`). It contains 361 paragraphs or 9,807 characters with so-called assignment operators for all letters of the Romanian alphabet, as well as for <k>, <q>, <w>, <x> and <y>, each separated into upper- and lowercase letters—the rules for upper- and lowercase are identical for the Latin and Cyrillic writing of Romanian. A very short readme file contains an explanatory description in English and Russian: "Script translits romanian (latin) to moldovan (cyr) characters Скрипт переводит румынские буквы в кириллицу. Можно переводить интернет- страницы. Например Википедию."[9]

Simple assignments can be made where a Romanian grapheme in Latin script has exactly one Cyrillic equivalent, e.g., <J> → <Ж> or <d> → <д>. Accordingly, the assignment operator consists of only one command (Fig. 4).

```
case "J":
   echo "Ж";
   break;
```

FIGURE 4. Script of a simple assignment operator

The graphemes of Romanian which have different phonetic realisations depending on their position—in particular <g> and <c>—must

---

9. Translation of the Russian sentences: The script translates Romanian letters into Cyrillic. You can translate websites. For example Wikipedia.

each be assigned to different Cyrillic letters, because the grapheme-phoneme-correspondences in Cyrillic Romanian or "Moldovan" are almost completely unambiguous.[10] Like in Italian, the Romanian <c> when preceding the vowels <e> and <i> is articulated as a voiceless prepalatal affricate [t͡ʃ]. This articulatory rule can be resolved by the digraph <ch> which, like <c> in all other positions, leads to velar articulation as [k]. In automatic transliteration, a complex assignment operator generates the letters <ч> and <к> depending on the position (Fig. 5).

```
case "c":
   {
   if ($str[$i+1]=="e" or $str[$i+1]=="i"
      or $str[$i+1]=="E" or $str[$i+1]=="I")
      { echo "ч"; break; }
   if ($str[$i+1]=="h" or $str[$i+1]=="H"
   and ($str[$i+2]=="e" or $str[$i+2]=="i"
   or $str[$i+1]=="E" or $str[$i+2]=="I"))
      { $i=$i+1; echo "к"; break; }
      echo "к";
      break;
   }
```

FIGURE 5. Script of a complex assignment operator

The following applies to complex assignments: Each larger defined unit of characters has priority over smaller character units or individual characters. So simple assignments can be supplemented by additional condition sets for special cases. Nevertheless, the grapheme <i> (or <I>), on which almost a quarter of the entire script is used, is at the limit of automatic assignment possibilities. As an isolated vowel, <i> should be rendered as <и>. However, especially at the end of a word, <i> is usually not a vowel but indicates palatalisation (<ь>, cf. Tab. 1, Ex. 2). In rising diphthongs, the iotified letters <ю> and <я> mentioned in Tab. 1, Ex. 5, are used. In falling diphthongs, such as <ei>, but also in the double <ii> that is frequent in Romanian, <й> is used, i.e., <ей> and <ий>.

A short text example from the Romanian Wikipedia (https://ro. wikipedia.org/wiki/Portal:Limbi) and the resulting transliteration will illustrate some of the complex assignments:

---

10. The only exceptions to graphophonic and phonographic unambiguity are the iotified letters <ю> and <я>, which are popular in Cyrillic (see Tab. 1, Ex. 5, and the transliteration of <i> below).

(1) O limbă reprezintă un sistem abstract, complex, de comunicare verbală în-
     tre oameni. În afară de forma orală (**limba vorbită**), bazată pe articularea de
     sunete, limbile actuale au în general și o formă grafică, **limba scrisă**.

(2) О лимбэ репрезинтэ ун систем абстракт, комплекс, де комуникаре
     вербалэ ынтре оамен**ь**. Ын афарэ де форма оралэ (**лимба ворбитэ**),
     базатэ пе артикуляр**я** де сунете, лимбиле актуале ау ын **ж**енерал ш**и**
     о формэ **г**рафикэ, лимба скрисэ.

There is nothing wrong in the automatic transliteration of the Ro-
manian text. Complex assignments are marked here, where the func-
tionality is clearly visible: the grapheme <g> is transliterated according
to the pronunciation once as <г> and once as <ж>. In the translitera-
tion of <i>, the assignment operator distinguishes between palatalisa-
tion in "оамень" and vowel realisation in "ши," whereby a final <i> is
in most cases to be read as a palatalisation sign, but not in monosyllabic
words such as *și*, where it must be vocal.[11] To end, in the transliteration
"articularea," it can be seen that the final vowels are not individually
recognised as hiatus, but as a diphthong, reproduced with <я>.

4.2.   Limits of the Automatic Transliteration

The text example shows overall that the assignment operators work.
However, in the following, some problems that cannot be solved with
the script are pointed out. Firstly, a number of special characters are
missing, i.e., letters with diacritics and other letters that go beyond the
basic Latin alphabet and that are not represented in Romanian. For
example, the name *Frédéric Chopin* is transliterated as *Фрѐдѐрик Кхопин*,
with the two vowels <é> remaining, and the digraph <Ch>, since it
is not placed before <e> or <i>, is interpreted as two single conso-
nants. Moldovan spelling would normally orientate on the Russian vari-
ant *Фридерик Шопен*. So, it is not only special characters that cause
difficulties, but also foreign-language sound patterns, which would be
rendered phonetically when transferred to another writing system. It
would be conceivable, however, to make automatic transliteration ca-
pable of learning for this purpose, since it follows the principle—as de-
scribed above—that each larger sequence of letters is given priority over
smaller combinations and individual characters. Accordingly, proper
names could be continuously included into the script.[12]

---

11. The final <i> is not recognisable in the script as a vowel in infinitives (e.g., *a veni*). This is discussed below (Tab. 5).

12. This alone, however, cannot solve the problem that when transliterating proper names, the original spelling would be added as a parenthesis. Particularly problematic is the retransliteration of proper names previously transcribed into the Latin alphabet, which are written in Cyrillic in the original.

Another problem are symbol letters and abbreviations. In chemistry, for example, element symbols are represented by Latin letters independently of the writing system of a language. In Wikimoldia, however, "H_2O" changes to "X_2O". For Roman numerals at least, a number of instructions has been preserved to prevent nonsensical transliteration. An incorrect abbreviation is pointed out in the forum for the script in *GitHub* (`https://github.com/slava37md2/wikimoldia/issues/2`): due to the positional phonetic realisation of the Romanian grapheme <c> (cf. Fig. 5), an error occurs in the abbreviation of 'centimetre' (Tab. 4).

TABLE 4. Incorrect automatic transliteration of Rom. *cm*

Romanian	correctly transliterated	automatically transliterated
cm	<чм>	<км>
km	<км>	<км>

Another special case of Romanian, which cannot be fully resolved in automatic transliteration, is homography, i.e., words that are spelt the same but pronounced differently. These include infinitives with the vocal ending *-i* (Tab. 5).

TABLE 5. Incorrect automatic transliteration of Rom. *dormi*

Romanian	pronunciation	correctly transliterated	automatically transliterated
tu dormi	[·dormʲ]	<дормь>	<дормь>
tu vei dormi	[dor·mi]	<дорми>	<дормь>

In principle, the palatalising function of <i> at the end of a word is much more frequent, so that the special case of the infinitive is less significant. Homographs sometimes occur between the infinitive and the conjugated form for second person singular in present tense. But numerous verbs of the *i*-group have stem extensions (e.g., *a citi—tu citești*) or irregularities (e.g., *a veni—tu vii*). In these cases, there are no homographs and the infinitives could be assigned to the correct transliteration as individual lexemes. For the remaining verbs with homograph forms (*a dormi*, *a fugi*, *a ieși*…) a fully automatic software should be able to distinguish between infinitives and conjugated verb forms. In view of the (Balkan-typical) restrictive use of the infinitive, which only allows its use in a few constructions with auxiliary verbs or the preposition *a*, it would be conceivable to program with so-called regular expressions,

which can formally map the syntactic embedding of infinitives beyond the level of lexemes.

The problems with automatic transliteration and their solutions presented in this section are purely technical. From a sociolinguistic perspective, however, the question may be asked concerning the interest of an error-free transliteration of Romanian into Cyrillic script. This will be the subject of the following section.

5.   The Potentials of Wikimoldia

The previous considerations about Wikimoldia are based on insights into the website when it was still active, on the analysis of the PHP script in GitHub, and on the analysis of usage data. The research for further background information on the creation and motivation of Wikimoldia remained fruitless. Thus, we can only speculate about the intended function of the website. The possible authorship, which is particularly important with regard to a conceivable political motivation, will also be discussed in this context.

Wikimoldia can be understood as an access to the Romanian Wikipedia for a Cyrillic socialised audience. This is particularly important for the population of Transnistria, but also for Romanian minorities in Ukraine. Although most Romanian speakers should be able to read the language in Latin script, in terms of literacy the possibility of reading in Cyrillic-script represents an additional benefit, provided that the readers do not switch completely to the Russian or Ukrainian Wikipedia. The above example, in which Wikimoldia was quoted on a Russian news site, also shows that Moldova's neighbours—potentially also Romania's Cyrillic-writing neighbours (Serbia, Bulgaria, Northern Macedonia)—benefit from the Cyrillic online presence as a bridge to Romanian-speaking culture.

Looking at the Wikipedia versions of different minority languages, it is often found that there is only a sparse number of rather short articles, so the value may be in the perception of the languages, but not in providing a useful encyclopaedia. In the case of Moldovan in the status described above, i.e., as a variety whose distance from standard Romanian is primarily defined by its script, the path of automatic transliteration provides the possibility of generating a comprehensive encyclopaedia on an *ad hoc* basis. While the previous section has explained how to improve the transliteration performance by computer, for a real encyclopaedia the suggestion could be made to freeze the transliterated version, to repair it manually, and to enrich it with individual content—although ethical and legal concerns may well be raised about the complete transfer of Wikipedia to Wikimoldia. Although Wikipedia content can theoretically be used freely as so-called open content and al-

though even the elaborately developed Wiki structure is freely available for the programming of independent variants, the transfer of the content without identifying the source is a plagiarism, and the replacement of *Wikipedia* by *Викимолдия* distorts the origin. But here as well, the genesis of the encyclopaedia could be made more transparent in a revision.

Finally, the question of the *cui bono* and thus the authorship of the Wikimoldia project can be raised. In the absence of precise indications, two quite contrary hypotheses on political motivation can be put forward: On the one hand, as a contribution to the vitalisation of the controversial Moldovan language, we could assume that the Transnistrian government and its post-Soviet continuation of the ideology of Moldovan "creat în laboratoarele Moscovei"[13] (Cimpoeșu and Musteață, 2018, p. 236) is the responsible agent behind this. On the other hand, the exact opposite can be assumed, namely Wikimoldia as a liberally oriented project that creates access to Romanian-influenced content and thus a rapprochement with the rest of Moldova, Romania and Europe, which, through the direct transmission of standard Romanian, ultimately even expresses itself linguistically in the aforementioned *limbă păsărească* that can also be understood as a protest against socialism (cf. Gabinskij, 2002, p. 135). These are, however, free speculations that would be conceivable alongside an apolitical interpretation of Wikimoldia, too.

6. Conclusion

In this paper, the curiosity of a Moldovan phantom page of the Romanian Wikipedia has been treated in its technical functioning and potential use spotting the sociolinguistic background of the Moldovan language and its controversial interpretation. The most striking element of Moldovan, and thus Wikimoldia, is the use of the Cyrillic alphabet according to the standard developed in the Soviet Union in the 1920s, which provides Romanian with an orthography that is largely harmonious with Russian, but phonographically particularly flat.

Since Wikimoldia is no longer online, we could dismiss as idle thoughts the pronounced ideas about improving and elaborating the project into a functioning encyclopaedia. In fact, the foundations that have been laid remain available in the form of the PHP script `slava37md2`, which could be used to reconstruct the site. However, because it has not become clear whether and what political ideology was behind Wikimoldia, and also because of the aforementioned concerns about the theft of data and ideas, the continuation and expansion of Wikimoldia should

---

13. Translation (with reference to Moldovan): created in the laboratories of Moscow.

perhaps rather be avoided. Irrespective of this, the programming of Wikimoldia can serve as a lesson for other multialphabetic languages, such as Serbian and other Serbo-Croatian or BCMS varieties, or even Hindi and Urdu. Within the Romance language family, Jewish Spanish is particularly worthy of mention, because it is maintained in its own Wikipedia (`https://lad.wikipedia.org/`) in two languages with the Latin and Hebrew alphabets, where the inequality in the expansion of the encyclopaedia is immediately apparent: There are much fewer and mostly only shorter articles written in Hebrew. The option of an automatic transliteration could also be a helpful support in this context.

References

Born, Joachim (2007). "Wikipedia. Darstellung und Chancen minoritärer romanischer Varietäten in einer virtuellen Enzyklopädie." In: *Sprachliche Diversität: Praktiken—Repräsentationen—Identitäten*. Ed. by Martin Doring, Dietmar Osthus, and Claudia Polzin-Haumann. Bonn: Romanistischer Verlag, pp. 173–189.

Cimpoeșu, Dorin and Sergiu Musteață (2018). *Basarabia la un secol de la Marea Unire. O istorie politică a Republicii Moldova. 1991–2018*. Târgoviște: Cetatea de Scaun.

Coulmas, Florian (2018). *An Introduction to Multilingualism. Language in a Changing World*. Oxford: Oxford University Press.

Dahmen, Wolfgang (2018). "Les frontières linguistiques extérieures du dacoroumain." In: *Manuel des frontières linguistiques dans la Romania*. Ed. by Christina Ossenkop and Otto Winkelmann. Berlin/Boston: De Gruyter, pp. 338–357.

Gabinskij, Mark A. (2002). "Moldawisch." In: *Enzyklopädie des europäischen Ostens*. Ed. by Miloš Okuka. Klagenfurt: Wieser, pp. 133–143.

Koch, Christian (forthcoming). "Wikimoldia – Викимолдия. Rekonstruktion eines Projekts zur Vitalisierung der moldauischen Sprache." In: *Korpus im Text* 10.

Kramer, Johannes (1989). "Rumänisch: Graphetik und Graphemik." In: *Lexikon der Romanistischen Linguistik*. Ed. by Günter Holtus, Michael Metzeltin, and Christian Schmitt. Vol. III. Tübingen: Niemeyer, pp. 14–18.

Leca-Tsiomis, Marie (2006). "Une tentative de conciliation entre ordre alphabétique et ordre encyclopédique." In: *Recherches sur Diderot et sur l'Encyclopédie* 40.41, pp. 55–66.

Olariu, Florin-Teodor (2017). *Variație și varietăți în limba română. Studii de dialectologie și sociolingvistică*. Iași: Institutul European.

Onu, Liviu (1989). "Rumänisch: Langue et écriture." In: *Lexikon der Romanistischen Linguistik*. Ed. by Günter Holtus, Michael Metzeltin, and Christian Schmitt. Vol. III. Tübingen: Niemeyer, pp. 305–324.

Zikmund, Hans (1996). "Transliteration." In: *Schrift und Schriftlichkeit. Ein interdisziplinäres Handbuch*. Ed. by Hartmut Günther and Herbert Ernst Wiegand. Vol. 2. Berlin: De Gruyter Mouton, pp. 1591–1604.

The Role of Punctuation in Translation

Dana Awad · Ghassan Mourad · Marie-Rose Elamil

Abstract. The objective of this paper is to address the problem of translating punctuation marks: is it really possible to translate punctuation the same way we translate vocabulary? or is it only a transfer of the functional use of punctuation marks from a source language to a target language? In order to study the role of punctuation marks in translation theories and practice, we first attempt to identify the nature and function of punctuation from a traductological perspective. The second part of our paper is a French-Arabic corpus analysis of the translation of punctuation marks. Our corpus analysis is based on Amin Maalouf's book *Les identités meurtrières* and its translation into Arabic.

Introduction

Punctuation marks have an important role in forming a logical sentence in order to communicate accurate meaning. Nevertheless, few studies highlight the importance of punctuation in the translation process. Previous research on the topic of punctuation in translation consist of contrastive linguistic analysis from a functional perspective and is targeted towards translation students. Those studies, though important, are descriptive, not based on a bilingual corpus analysis to support their findings and focus on the grammatical aspect of punctuation, hence discarding the cognitive nature of those marks.

In this paper, we suggest different methods to translate punctuation marks based on translation theories and we observe the translation of

Dana Awad   ⓘD 0000-0001-9522-4549
Lebanese University, al-Kantari, Moutran Shebli Str., Beirut, Lebanon
E-mail: `dana.awad@ul.edu.lb`

Ghassan Mourad
Lebanese University, al-Kantari, Moutran Shebli Str., Beirut, Lebanon
E-mail: `ghasmrad@hotmail.fr`

Marie-Rose Elamil
Lebanese University, rue principale Al Marej, Rmeich, Nabatieh, Lebanon
E-mail: `marieroseamil38@gmail.com`

punctuation by a professional translator based on a French-Arabic corpus of Amin Maalouf's *Les identités meurtrières* and its translation into Arabic. Our paper aims to highlight the important role of translating punctuation marks not only in sentence structure, but also in communicating the tacit meanings intended by the author.

Related Works

Previous research on the importance of punctuation in translation either have a didactic approach centered towards translation students or a contrastive linguistic approach; we have failed to find a research paper that studies punctuation from a traductological perspective.

The didactic approach offers a method to train students to translate punctuation by relating it to the grammatical system of the source and target languages, and then to create benchmarks of the main differences. Spilka (1988) categorized those differences between French and English into the following:

- orthographic differences, such as the difference between English and French quotation marks;
- typographic differences, such as spaces before or after a punctuation mark;
- syntactic differences (Spilka gives the example of different French and English enumerations and adding punctuation between connectors, but this can also include differences in organizing independent and dependent clauses);
- textual differences[1].

Another didactic approach study is written by Mogahed (2012). In his paper, Mogahed tries to show the importance of punctuation in keeping or changing the meaning of the source text by using examples of English-Arabic translations. For example, he points out how the meaning changes in the following sentence when a semicolon is used instead of a comma (Mogahed, 2012, p. 3):

> I have taken several science courses this year; my favorite was neuroscience.
> I have taken several science courses this year, but my favorite was neuroscience.

Mogahed explains that joining the two independent clauses with a comma and a conjunction (but)

---

1. By textual differences, Spilka refers to paragraph segmentation and the use of the typographic mark *alinea* to indicate a new paragraph. Although the topic of paragraph segmentation is, like punctuation, related to the organization and cohesion of ideas, we will focus in this paper on punctuation marks as linguistic signs.

changes the meaning slightly from the previous version; it emphasizes the contrast between the group of courses in the first clause and the single course in the second clause. (ibid.)

Although these studies are important, their interest is specifically for translation students and the examples they use are not based on a corpus[2]. Nevertheless, those studies make an important point: punctuation marks, in translation teaching and practice, are sometimes unstated. The reason for this might be that they are considered universal, so even if the standard use of punctuation is not respected, the TL reader will be able to decipher the ST meaning. This is also because general rules of punctuation (at the syntactic level) are almost similar in all languages: a comma always coordinates words and dependent clauses; a semicolon always coordinates independent clauses, etc.

Research based on contrastive linguistics (Alqinai, 2013; Ponge, 2011) focus on the importance of linguistic analysis of punctuation and on comparing the relationship between punctuation and sentence structure in a source and a target language. When both language systems are very distinct, orthographic marks are also added, for example, Alqinai mentions the absence of capitalization in Arabic and how, when necessary, capitalization is translated into round brackets or quotation marks. Ponge (2010) bases her linguistic analysis of punctuation on Nina Catach's work, and although her work is about punctuation in translation, most of her research paper is a about the nature of punctuation from a linguistic point of view. According to Ponge, who is inspired by N. Catach's work, punctuation is similar to ideograms that have an international connotation, which makes them "*semantically stable*" (Ponge, 2011, p. 124), and therefore, a creation of benchmarks for translators is easy.

Since contrastive linguistics is closely related to translation studies, such research work is important to highlight the relationship between punctuation and target text readability and organization of information. Although some translation techniques are mentioned, like back translation to insure that punctuation did not change the meaning in the source text, previous contrastive linguistics centered work on the subject fail to properly place punctuation within translation theories.

In our paper, we attempt to fill this aforementioned gap by attempting to integrate punctuation into existing translation theories, and we will demonstrate the role of punctuation in translation practice via a corpus analysis of Amin Maalouf's book *Les identitiés meurtrières* and its Arabic translation. Based on this analysis, we will see if the translation of punctuation is only functional[3], due to syntactic differences between both languages, or if there is a semantic translation to punctuation.

----

2. The source of the examples is not mentioned in those studies.

3. What we call functional translation is the translation process of grammatical entities that do not have a meaning, but a function like tenses and pronouns.

In order to relate punctuation marks to translation theories, we searched for theories that focus on the form of the source and target text and its relation to meaning rather than theories that focus on equivalence[4] or on the question of fidelity in translation. We specifically attempt to give a traductological approach to the role of punctuation in translation through the works of Eugine Nida's *Toward a science of translating* (1964) and Marianne Lederer's interpretive method of translation (1994). In order to integrate punctuation to these theories, we find it necessary to categorize[5] the different roles of punctuation marks[6]:

1. Syntactic punctuation marks that organize knowledge in sentences and paragraphs and that hierarchize information, making some information essential to communicate in the text and other information less important or only necessary to clarify or enhance the essential information. These punctuation marks are the period, comma, double commas, semicolon, colon, parenthesis and double dashes.
2. Semantic punctuation marks that add a connotation that can only be communicated by punctuation such as quotation marks, question mark, exclamation point and ellipsis.
3. Punctuation marks that give reference to another author or speaker such as quotation marks, colon and dash.

Form and Meaning in Translation Theories

Nida wrote his book *Toward a science of translating* to answer the dilemma translators face when translating literary work, which is whether to keep the translation literal even if this would result in a target text stylistically foreign to the target language, or to translate the meaning while changing the form to a stylistically acceptable one in the target language. This dilemma, which he calls "*the letter vs. the spirit*" (Nida, 1964, p. 3) is due to the fact that languages are very different; while translating literally is a kind of "protection" for the translator as a proof of fidelity to the text, it can hinder the communicative purpose of translation. In order to solve this dilemma, Nida referred to translation as a science rather than an art, providing a scientific description of the process of transmitting a

---

4. One of the results of our corpus analysis is that change of punctuation in the target text is often the result of the translator's choice to change the form of the source text. Nevertheless, we found examples of punctuation translated by a functional equivalence.

5. This is a general categorization in order to integrate punctuation into translation theories; we suggest a slightly different categorization in our corpus analysis.

6. Due to the multi functionality of each punctuation mark, some punctuation marks exist in more than one category, and, if one conducts a monolingual corpus analysis of punctuation, all punctuation marks can exist in all categories in some languages, such as Arabic (Awad, 2013).

message from one language to another. By considering translation a science, we separate translation from the art of writing and in that way, when a translator chooses to change the form of the source text, he is not writing a different literary text[7] but adapting a scientific process of translation. Nida sees translation as a practice that can be divided into three types:

1. Intralinguistic: within the same language using different words.
2. Interlinguistic: or, as he calls it, translation proper, which is the transmission of verbal signs of a language by the verbal signs of another.
3. Intersemiotic: translating from one system of symbols to another.

Nida's interlinguistic translation process incorporates Chomsky's transformational grammar theory in order to produce a "science" of the translation process (Munday, 2008, p. 57), thus turning the practice of translating into a rigorous process of transferring knowledge and to organizing this knowledge in an accepted way in the TL. Nida's science of interlinguistic translation is a process of decoding the source text and recoding the target text: a sentence or a paragraph in the source text is called, by reference to its form, a *surface structure.* When translating, the surface structure is decoded in order to understand its *deep structure* (the message it conveys), and the translator recodes this deep structure into another surface structure in the target language. In the final surface structure, the translator has to choose words and form according to the target language system in order to create a meaningful and readable content for the target audience[8].

Source sentence —— The message it conveys —— Target sentence

SL surface structure ——— Deep structure ——— TL surface structure

---
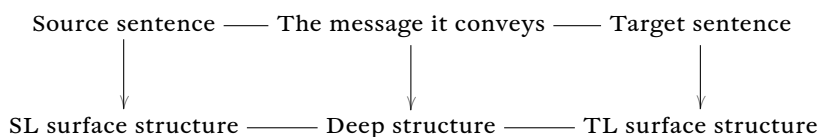
7. In the introduction of his book, Nida cites incidents where authors and reviewers of translations were reluctant to the communicative approach of translations for not being loyal to the stylistic aspects of the source text. He gives an example of translations being criticized or rejected because the form of the target text was very different from the source text, which was "a descriptive essay of rhythmical sentences, simple phrases and well-chosen words." (Nida, 1964, p. 1), which, in result, made a back translation impossible.

8. This theory is similar to Catford's notion of "translation shifts" (Catford, 1965, p. 20), in (Munday, 2008, p. 95) and achieving a textual equivalence through different categories of shifts including structural shifts (syntactic structure) and rank shifts that refer to linguistic units such as clause, sentence etc. Although, like all other translation theories, punctuation is not mentioned, but we think that syntactic punctuation is implied as included in the "structural shifts".

We find that Nida's science of translation can be adapted in the process of translating punctuation. Creating a TL appropriate surface structure includes the formal aspects of punctuation, such as different spacing before and after a punctuation mark, and syntactic differences between two languages. An example of syntactic differences is the relationship between punctuation marks and connectors: using punctuation before, after or instead of a connector. Translating deep structure of punctuation focuses on the semantic properties of punctuation and on keeping the same hierarchy of knowledge of the ST deep structure. By applying Nida's science of translation, translators will focus on the deep structure of the source sentence and then recode it into a surface structure appropriate to the target language system instead of having a list of functions for punctuation marks and their equivalences in another language[9].

The connotative nature of punctuation marks adds a meaning that cannot be transmitted by words or gives connotation of the importance of information. Translators, therefore, need to acquire the skill of interpreting the use of punctuation in the source text: is punctuation used by the ST author an "auxiliary" for grammatical and syntactic purposes, a necessary tool for semantic connotation or a stylistic tool that transmits the author's individuality and spirit? We attempt to adapt this interpretative process to Marianne Lederer's interpretative approach to translation.

The interpretative approach to translation takes into account all compositions of a text: the linguistic elements as well as the extra linguistic elements that belong to printed characters (layout). The latter should be distinguished from linguistic elements because, according to Lederer, layout of the printed text is not taught in schools as language is[10]. Therefore, the translator must separate sentence structure and text structure when translating. The translation process includes three levels: language (words), sentence construction and text[11].

Although the different stages of the translation process in the interpretative method resembles Nida's stages (understanding, deverbalization and re-expression), Lederer is more focused on the cognitive aspect of translation. The first stage in the translation process is understating

---

9. It is, however, important for translators to have a list of equivalences for non-existent typographic marks (in the case of Arabic, capital letters and italics) and replacing them with punctuation when needed.

10. On this point, we disagree with Lederer because some layout elements, such as paragraph separation, is taught in schools. However, even though paragraph separation rules are universal (separation of paragraphs is related to separation or linking distinct ideas), we can notice that the number of paragraphs slightly differs when translating from one language to another. This difference can be explained by the separation of linguistic and extra linguistic elements in the translation process.

11. Lederer developed this process for interpretation and for written translation.

through reading the entire text and analyzing the intentions of the author before re-expressing those intentions in a TL. This includes vocabulary, segmentation, and punctuation (though the latter is not specifically mentioned by Lederer).

By applying Lederer's theory to the translation of punctuation marks, they will be considered as not only signs associated with a function, but also as holders of *cognitive complements* that are constructed in the language system and that are part of the reader's background knowledge. For example, in an Arabic reader's background knowledge, a double dash is used for parenthetical elements repeatedly written after the name of a religious or an authoritative figure. Therefore, the element between double dashes will be discarded or read with less attention. The translator would therefore translate the double dash into Arabic with another punctuation mark that holds the same cognition of the double dash in the source text. Another example is the colon that is rarely used in Arabic texts[12], unlike English or French where a colon is used to coordinate clauses.

Contrastive Analysis of Punctuation and Its Role in Translation

In order to conduct a contrastive analysis of the use of punctuation marks between a source and a target text, We would categorize punctuation marks into the following:

1. Punctuation for syntactic purposes: this includes punctuation necessary for sentence structure (comma, semicolon, period) and the order of clauses. In this case, a contrastive analysis of punctuation is related to contrastive grammar. For example, how independent and dependent clauses are separated differently in two languages, how words are connected in case of enumeration (when to use a comma and when to use a semicolon, for example), and the relationship between the comma and connectors in a language. Although this contrastive analysis seems intuitive for translators, who have full master of source and target languages, questions regarding punctuation seem to occur often. For example, whether to put a space before or after a punctuation mark. Another question regarding the structural purposes is the attachment, or not, of a comma with the Arabic connectors *wāw* and *fāʾ*. Since both convey several meanings, the comma determines the meaning of those polysemic connectors. For example, when using a comma before *fāʾ*, the particle in this case conveys a cause/result relation, and when we don't use a comma before *fāʾ*, it

---

12. In general, the colon is used in Arabic academic texts, dialogues (after the name of the speaker), and in newspaper titles.

indicates the chronological order in which verbs occurred. We can divide punctuation marks in this category into:

- Punctuation that divides clauses;
- Punctuation marks that are used with a connector;
- Punctuation marks that act as connectors: like the colon and the semicolon, for example.

For syntactic punctuation, literal translation is possible if both languages have similar grammatical systems (for example, Italian and French). In general, there is a degree of universality to the grammatical functions: punctuation marks are either separators of clauses or of sentences. The question of translating syntactic punctuation therefore lies in the language's preference in relating clauses using punctuation, another connector, or both, and in the different structure of clauses between both languages. Changes in placement of punctuation marks is, therefore, dependent on the target language and structuring sentences in a "native"[13] way.

2. Punctuation for hierarchization of information: to hierarchize information, double punctuation is used: a double comma, parenthesis, quotation marks for sentences, double dashes. In regards to the hierarchy of knowledge, punctuation marks have different connotations in different languages. For example, double dashes connote unimportance of information in Arabic, while in French the information would be more important if we use dashes instead of parenthesis. Extra linguistic knowledge of the role of punctuation in highlighting or marginalizing information is necessary to understand the writer's intentions and to translate them accordingly. Regarding this role, Literal translation of punctuation is not possible because each punctuation mark hierarchizes a text differently in different languages.

3. Punctuation for expressing emotions and intentions: like the exclamation point, question mark, ellipsis and quotation marks for words. In practice, translators often choose to keep punctuation for emotions and intentions as they are in the source text because they are cautious not to "over analyze" or over interpret them, even if their connotation is expressed by another punctuation mark in the target language[14].

---

13. Changing syntactic punctuation of the source text is for the target text to become "natural" according to the target language syntactic rules so that the reader does not feel that the text is a translation. Most of the time, the target text is readable and understandable whether syntactic punctuation is translated literally or not, but the reader can notice the foreignness in language structure.

14. For example, irony in Arabic is often expressed in Arabic literal and journalistic texts by an ellipsis before the word and an exclamation point after it; however, translations usually keep the ironic quotation marks when used in the source text.

When translators analyze the components of a source text, syntactic and hierarchical punctuation marks are considered as part of the written language system. On the other hand, punctuation marks for emotions and intentions are dependent on the author's style. Therefore, translators try to keep them as they are to protect the authenticity and the spirit of the writer who, for translators, uses punctuation marks for emotions and intentions as tacit marks that are understandable in the text, but what is understood should not be translated.

Corpus Analysis

Our French-Arabic corpus analysis is of Amin's Maalouf's book *Les identités meurtrières* (Maalouf, 1998) and its Arabic translation (Maalouf, 2011). Our analysis process was to collect punctuation from the source and target text, classify them according to the aforementioned categories and analyze the translator's behavior in translating punctuation and why did she choose to punctuate her text differently.

After collection of punctuation in the ST and the TT, the result was that the translator used less punctuation than in the original version:

Punctuation mark	Number of occurrences in the source text	Number of occurrences in the target text
Period	1,097	911
Comma	3,784	2,415
Semicolon	242	86
Exclamation point	25	24
Question mark	164	163
Colon	77	70
Double dashes	159	132
ellipsis	36	42
Quotation marks	275	275
italics	6	0

From the table, we see that Amin Maalouf's punctuation for this book is classified as follows:

1. Syntactic punctuation: period, comma, semicolon;
2. Punctuation for hierarchization of information: quotation marks, double dashes, and italics (as a typographical mark);
3. Punctuation for emotions or intentions: question mark[15], exclamation point, ellipsis, quotation marks.

---

15. The question mark can also be syntactic if its use is to mark the end of a question. In this paper, we categorized the interrogative nature of the question mark as "intention".

From the previous table, we can say that the translator chose not to translate some punctuation marks and that she chose to use connectors instead of punctuation in some sentences. We will show examples of sentences that were punctuated differently in the translation.

Examples From the Corpus

Quotation Marks

In the following example, both quotation marks and the words between quotation marks are omitted in the translation. The author intended to add a sarcastic connotation to the expression by using quotation marks. The translator, however, chose to omit the full segment instead of translating it. Our hypothesis is that since the segment is part of a long list of enumeration of the physical aspects of a person, the translator felt that removing a part of this enumeration would not affect the general meaning of the sentence.

فالآخرون يُشعرونه، بكلامهم وبنظراتهم أنه
فقير، أو أظلع، أو أسمر البشرة، أو شديد البياض،
أو مختوناً، أو غير مختون، أو يتيما إنّ كل هذه
الإختلافات العديدة، الثانوية منها والجوهرية، هي
التي ترسم ملامح كل شخصية،

Les autres lui font sentir, par leurs paroles, par leurs regards, qu'il est pauvre, ou boiteux, ou petit de taille, ou "**haut-sur-pattes**", ou basané, ou trop blond, ou circoncis, ou non circoncis, ou orphelin—ces innombrables différences, minimes ou majeures, qui tracent le contours de chaque personnalité

In the following example, the translator did not consider the importance of highlighting the word between quotation marks to draw the reader's attention. By omitting the quotation marks, the author's intention was not accomplished. The bold character[16] would be an equivalence for quotation marks in this case.

فإذاعاش المرء في بلد يخشى فيه الإعتراف
بأن اسمه بيير أو محمود أو باروخ، وكان هذا
الوضع مستمراً منذ أربعة أجيال أو أربعين جيلا
، إذّا عاش المرء في بلد لا يحتاج فيه إلى
مثل هذا **الإعتراف** لأنه يحمل أصلاً على وجهه

Si l'on vit dans un pays où l'on a peur d'avouer qu'on se nomme Pierre, ou Mahmoud, ou Baruch, et que cela dure depuis quatre générations, ou quarante; si l'on vit dans un pays où l'on n'a même pas de faire un tel "**aveu**", parce qu'on porte

---

16. We added the bold character both in the French text and in the Arabic translation of the example.

déjà sur son visage la couleur de son appartenance, parce qu'on fait partie de ceux qu'on appelle dans certaines contrées "les minorités visibles"; alors on n'a pas besoin de longues explications pour comprendre que les mots de "majorité" et de "minorité" n'appartiennent pas toujours au vocabulaire de la démocratie.

لون انتمائه، ولأنه ينتمي إلى أولئك الذين يعرفون في بعض الدول «بالأقليات المرئية»، فلا حاجة عندئذٍ لتفسيرات مطولة كي نفهم أن مفردات على غرار «أكثرية» و «أقلية» لا تندرج دوما في قاموس الديمقراطية.

The Comma

In the following example, the author used a double comma as a form of hierarchizing information: what is between double commas has a double role in the French sentence: it highlights the discriminating nature of the question regarding his origins and, more importantly, it expresses the feelings of the author, who is not offended by the question. Although the main meaning is the same in the Arabic sentence, the role of the double comma was not translated.

[...] que de fois m'a-t-on demandé, avec les meilleures intentions du monde, si je me sentais "plutôt français" ou " plutôt libanais"

فكم من مرّة سألني البعض عن طيب نيّة إن كنت أشعر بنفسي «فرنسيا» أم «لبنانيا».

The comma is often translated by the particle *wāw*, especially in enumeration.

Qu'il s'agisse de la langue, des croyances, du mode de vie, des relations familiales, des goûts artistiques ou culinaires, les influences françaises, européennes, occidentales se mêlent en lui à des influences arabes, berbères, africaine, [...]

سواء تعلّق الأمر باللغة والمعتقدات وأسلوب العيش والروابط الأسرية والأذواق الفنية أو أنواع المأكل لأن التأثيرات الفرنسية والأوروبية والغربية تمتزج في كيانه بالتأثيرات العربية والبربرية والأفريقية والإسلامية [...]

The Colon

As we mentioned previously, the use of a colon gives an academic or journalistic aspect to Arabic text. Therefore, the translator uses connectors that convey the same meaning of the colon. In the first example, the translator translated the colon by the particle *kāf* (used before examples), and in the second example, she translated the colon by *wa hiya taqūm ʿalä* (it is based on).

Puis des idées nouvelles ont lentement réussi à s'imposer : l'idée que tout homme avait des droits qu'il fallait définir et respecter	ثم بدأت أفكار جديدة تفرض نفسها شيئاً فشيئاً كالفكرة القائلة إن كل إنسان يتمتع بحقوق يجب تحديدها وإحترامها،
[...] il y a un jeu mental éminemment révélateur : imaginer un nourrisson que l'on retirerait de son milieu à l'instant même de sa naissance pour le placer dans un environnement différent	[...] ثمّة لعبة ذهنية معبَّرة للغاية، وهي تقوم على تصور طفل رضيع فصل عن محيطه منذ ولادته،ونقل عن بيئة مختلفة عن بيئته الأصلية،

The Period

The translator often chose to omit the period in order to connect two sentences instead in the original text for syntactic reasons. In the first example, the two sentences in the source text are semantically related to each other because the second sentences is an example of *éléments* in the first sentence. The preferred structure in Arabic would be to include the examples in the same sentence. In the second example, the second sentence starts with *mais* (but), and in Arabic, it is grammatically incorrect to use a period before *mais*, so keeping the period in the translation is impossible in this case.

[...] une foule d'éléments qui ne se limitent évidemment pas à ceux qui figurent, sur les registres officiels. Il y a, [...]	[...] جملة عناصر لا تقتصر بدهيا فحسب على تلك الواردة في السجلات الرسمية،ومن بينها، [...]
Le bon sens voudrait qu'il puisse revendiquer pleinement cette double appartenance. Mais rien dans les lois [...]	ويفترض المنطق السليم أن يستطيع هذا الشخص المجاهرة بانتمائه المزدوج،ولكن لا شيء في القوانين [...]

Conclusion

Although translation theories mention textual equivalence to explain change of form and structure, they do not mention punctuation as an important element to achieve this structure. We believe the reason for this omission is either because translation theorists think it is a given addition not necessary to mention—because its inclusion is implied when we say text or structure—, or because of the highly interpretative nature of punctuation that might put the fidelity to the source text at risk if translated. In other words, if punctuation is overanalyzed, the process

might hinder the transmission of meaning since these extralinguistic units are highly interpretative both at the level of hierarchization of knowledge and of adding information or "emotions". This explains why, according to our findings, translators prefer to omit punctuation if it does not have a syntactic function in the original text and only translate the main meaning of a sentence. In doing so, their translation is incomplete because they exclude tacit information that the author expressed using punctuation marks, especially if it is a literary text. This omission also proves that punctuation marks do not have universal meanings or functions. In order to translate punctuation marks, they must be regarded as signs with a semantic capacity equal to words, and therefore, should be divided into different semantic categories that are specific to a language.

References

Alqinai, Jamal (2013). In: *Linguistica Atlantica* 32, pp. 2–20.

Awad, Dana (2013). "La ponctuation arabe: histoire et règles, étude contrastive arabe, français, anglais." PhD thesis. Université Lumière Lyon II.

Catford, John C. (1965). *A Linguistic Theory of Translation*. Oxford: Oxford University Press.

Jungwha, Choi (2003). "The Interpretive Theory of Translation and Its Current Applications." In: *Interpretation Studies* 3, pp. 1–15.

Lederer, M. (1994). *La traduction aujourd'hui*. Paris: Hachette.

Maalouf, A. (1998). *Les identités meurtrières*. Paris: Grasset.

———— (2011). الهويّات القاتلة [*Killer Identities*]. Trans. by Nahla Baydoun [Dār al-Fārābī] دار الفارابي [Beyrouth:] بيروت. [نهلة يضون].

Mogahed, M. (2012). "Punctuation Marks Make a Difference in Translation: Practical Examples." Faculty of Education, Mansoura University Curriculum & Instruction Department, https//files.eric.ed.gov/fulltext/ED533736.pdf.

Munday, J. (2008). *Introducing translation studies: theories and applications*. London: Routledge.

Nida, E. (1964). *Toward a science of translating*. Leiden: Brill.

Ponge, M. (2010). "Signes de ponctuation: indices de traduction (Analyse comparée français / espagnol)." In: *Actes du colloque "Traduction, changement en syntaxe, personne – Approches fonctionnalistes", Oct 2010, Corfou, Grèce*, pp. 69–72.

———— (2011). "Pertinence linguistique de la ponctuation en traduction (français-espagnol)." In: *La linguistique* 47, pp. 121–136.

Spilka, I.V. (1988). "Comparer pour traduire." In: *Méta: Journal des traducteurs / Meta: Translator's journal* 33.2, pp. 174–182.

Comparing the Visual Untranslatability of Ancient Egyptian and Arabic Writing Systems

Hany Rashwan

Abstract. The paper discusses the mechanism of visual *jinās* (roughly meaning wordplay?) in both ancient Egyptian and Arabic languages. It demonstrates the significance of looking into several overlooked visual aesthetics, which were mainly designed to stimulate the eyes and minds of the indigenous readers, to shape any theory related to the literary analysis of ancient Egyptian or Arabic writing systems.

Viewed linguistically, the AE language belongs to the same phylum as Arabic, which is known now as Afroasiatic[1] and shares many of the same linguistic features fundamental to literary production. Comparative linguistics has been concerned from the early nineteenth century with describing and arranging African-Asian languages and generating linguistic theories about their historical development, especially after deciphering the Ancient Egyptian and many other ancient languages. (Hodge, 1983) The Afro-Asiatic phylum has a history of scholarship acceptance almost as long as that of Indo-European, despite being a fam-

---

Hany Rashwan   (iD) 0000-0001-6963-6603

University of Birmingham, School of Languages, Cultures, Art History and Music. Edgbaston, Birmingham, B15 2TT, UK

E-mail: H.Rashwan@bham.ac.uk

1. The term Afroasiatic is also known as Afrasian (Diakonoff, 1981), Hamito-Semitic, Semito-Hamitic. The biblical terms such as Hamitic, Semitic, and Cushitic led to the long use of Hamito-Semitic or Semito-Hamitic for the whole phylum. Nowadays, these terms are objectionable because of their mythological origins; thus a neutral geographic term "Afro-Asiatic" or "Afro-asiatic" was generated and came into usage. The old opposition of Semitic to a certain "Hamitic" unity (into which all the African members of the family were forced) was resolved in the 1950s by Greenberg. He argued for the equal status of four African branches beside the Egyptian: Berber, Chusgitic, Omotic, Chadic.

ily of much greater internal diversity and historical time depth. (Ehret, 1995)

This proposed comparison entails a conscious rejection of imposing Eurocentric concepts and terms — according to which Euro-American researchers did not support their literary assumptions by comparing AE literary devices with those of its kindred languages. Instead, I have sought to look outside literary studies altogether, and to apply the main principle of the comparative linguistic system: "languages should never be compared in isolation if closer relatives are at hand." (Greenberg, 1971, pp. 22–23) This statement is particularly relevant when dealing with a 'dead' language. Studying the AE language is archaeology of a dead language, in which cross-linguistic comparisons provide strong support for closer hypotheses on literary textual practices, to avoid Eurocentric rhetorical misperceptions and misrepresentation. Stephen Quirke looks to the use of Arabic linguistic affinities with their AE counterparts. He explains how their interaction with the Arabic literary tradition could be useful for both AE literary analysis and for challenging Eurocentrism in the field of Egyptology as a whole. He argues that European-language impositions will not fully resolve the problematic questions raised by AE literature. Therefore, Quirke encourages Euro-American scholars to give the Arabic literary world a chance equal to the one that has been offered to their Eurocentric theories. Quirke considers that active engagement with Arabic literary traditions promises fresh perspectives that may challenge the self-contained approaches of contemporary theoretical readings of AE texts:

> Classical Arabic poetry offers for certain motifs and 'genres' a resonance entirely lacking in English and other European literary traditions. The eulogy genre *madiḥ* allows appreciation of compositions at or outside our literary borders, and the *fakhr* 'boast' mercifully loses in Arabic the unfailingly negative reception assigned to much rhetorical content in English language studies of both literary manuscript and 'autobiographical' inscriptions from ancient Egypt. A more systematic encounter with Arabic literary tradition would above all serve to remind the European researcher that the questions of definitions, production, and reception of ancient Egyptian literature can also be asked from within Egypt. (Quirke, 2004, p. 28)

The comparison offered is part of a new suggested discipline called Comparative *Balāghah*. This new discipline focuses on comparing the literary devices of two kindred languages productively. I mean by 'productive' that the stylistic differences between the two systems are more stressed than the similarities. (Rashwan, 2020, p. 391) This methodology argues that AE literary devices are studied most productively on a comparative basis and that Arabic, a cognate language that belongs to the same Afro-Asiatic phylum, offers a new and closer platform for exploring and studying these literary devices. The literary structure of every language is peculiar to itself. The logic of this comparison argues

that a close investigation of the literary worlds of both the premodern Arabic and AE can shed new light on the literariness of the AE writings, by highlighting the importance of rediscovering the various forms of each literary device and their semantic function inside the studied text.

The comparative reading of these AE literary devices offers the occasion for one further point of argument, and that has to do with how scholars should approach the literariness of the AE texts more broadly, opening the door to previously unexplored literary and linguistic approaches. Comparative *balāghah* keeps the conversation and literary engagement going. It extends the conversation, opens it out, and makes it potentially relevant to issues and interests not foreseen at the outset. Using this comparative methodology will not only supply exceptionally innovative insights into how the AE language makes literature, but it can achieve the required depth and complexity to answer many new questions that are not even envisaged by the Western literary traditions.

Using the Arabic frame to rediscover the AE literary practice does not imply forcing the investigated materials into Arabocentric concepts and definitions. Comparative balāghah aims to understand the native term first and see how it is similar or different from the Arabic and Eurocentric ones, to find a shared platform that may develop our conceptual understanding about what can be accepted as universal or neutral terms. The differences between the two languages are more important than the similarities; this point is well-illustrated in the visual *jinās* study of the AE writing, which rediscovers the ability of the AE writers to build visual metaphors that are carried by clever employment of the soundless determinatives that visually reflect the verbal layer.

Visual *jinās* (التصحيف—المرسوم—الخط)

This term in Arabic refers to two words that have identical number and kinds of letters except for one different letter, and these two letters are graphically similar (Al-Gundy, 1954, p. 140), such as (س—ش), (ر—ز), (ص—ض), (ط—ظ), (ع—غ), (ج—ح—خ), (ي—ب—ت—ث—ن), (د—ذ), (ف—ق). The writer Mohammad ibn Badr el-dīn al-bisāṭī (d. 1634) collects various examples of this type of *jinās*, in his book entitled, *The inherited and humorous in the art of visual jinas* (*al-tālid wa al-ṭārīf fī fann jinās al-taṣḥīf*). Al-bisāṭī considers it one of the most wonderful types of *jinās*, because the mind of the ardent reader is being stimulated by considering the visual and verbal interactions that are rooted in the nature of Arabic script, whether in poetry (*manzūm*) or literary prose (*manthūr*) (Al-Bisāṭī, 2018, p. 32).

وَالَّذِي هُوَ يُطْعِمُنِي وَيَسْقِينِ وَإِذَا مَرِضْتُ فَهُوَ يَشْفِينِ

(God) He is the one who feeds and waters me, and when I become ill, He is the one who cures me. Q26:79-80

Visual *jinās* is represented between the two words يسقين—a verb meaning 'to make me drink water' and يشفين—a verb meaning 'to cure my health problem'. Each word contains two graphically similar letters (س—ش) and (ف—ق).

The open question raised here is—can we apply the Arabic understanding of this visual type of *jinās* to the AE writing? i.e., the AE signs that look alike graphically[2], for example, the bird pictures 𓅿𓅿𓅾𓅽𓅬𓅢𓅧𓅩 or human figures 𓀀𓀁𓀂𓀃𓀄𓀅𓀆𓀇𓀈. I would argue that the main difference between AE writing and any other alphabetical system would be related to how the AE writers take advantage of the visual inimitability of their writing. The examples provided in this section shows how the Egyptian writer can supply the reader with various visual tools for better understanding the points he raises about the presentation and structure of information, in order to aid and clarify the visual reading process. The examples cited here for visual *jinās* confirm that visual clarification of the intended meaning is believed to be one fundamental function of the AE determinatives. A more careful analysis of the neglected role of the determinatives inside the AE sentences reveals that their functions are much more versatile literarily. The AE writer built literary texts out of words, but every word, if carefully examined in its textual context, will turn out to be a literary volcano in itself.

Related Determinatives

The AE writer employs related 'determinatives' or 'sense-signs' as images that reinforce the sequence of his words to build a visual context that confirms for the reader's eye the verbal message, eloquently.

𓏃𓈗𓈖𓂝𓏭𓈗𓏃𓂝

mi mw ḥm ibt

Like the water when it quenches the thirst. (Eloquent Peasant, Parkinson, 1991, pp. 34, l. 278)

---

2. On cryptography (the use of hieroglyphic signs to denote sounds or meanings different to their usual use often involves groups of similar signs), Darnell (2004, p. 3) argues in the introduction to his book entitled The Enigmatic Netherworld Books of the Solar-Osirian Unity: "At no time, however, do the Egyptians appear to have considered hieroglyphic cryptography as something other than an extension of the normal hieroglyphic system, for they do not appear to have employed any separate term for "cryptography". He cites the most highly developed play on similar signs, in this mode of "cryptography", in the temple of Esna (Roman Period) where one hymn is written with variations of ram signs and another hymn has crocodile signs—see photographs in Hallof (2011, p. 10).

Visual *jinās* is represented between three related sense-signs that visually stress the meaning of being thirsty: 〰️—*mw*—a noun meaning 'water'; 𓐍𓅓 *ẖm*—a verb meaning 'to quench (thirst) with just water as an ending determinative 〰️; and 𓇋𓃀𓏏 *ibt*—an infinitive form of the verb 𓇋𓃀 *ibi* that means 'being thirsty', with an ending determinative of a man putting his hand towards his mouth, showing his need for water, and with a leaping calf, as a thirsty creature might jump at a source of water. The repetition of the water's sign 〰️ as an ideographic sign in the word *mw* and as a soundless ending determinative in the verb *ẖm* highlights its importance for the following word 𓇋𓃀𓏏 *ibt*, whose determinatives represent two thirsty figures from the human and animal worlds (〰️—〰️—🐾—𓃒).

first word	second word	third word	related images
〰️〰️	𓐍𓅓〰️	𓇋𓃀𓏏	〰️ - 〰️ - 🐾 - 𓃒

𓇋𓅓𓏤𓈖𓏏𓅱𓈖𓇋𓐍𓏏𓆑𓐙�Ꮠ�𓏺𓃀𓐍𓆑𓇋𓐙𓏤𓅂𓄣𓆑𓅓

im in.tw n.i ḫꜣt.f mꜣi pḫfy mꜣi ḥr-ib.f m

Let someone bring to me one whose front is a lion and his back is a lion and his middle like... (Khakheperreseneb text; Parkinson, 1997, p. 64)

Visual *jinās* is represented by the use of two successive words whose meanings are related to each other in a stimulating visual way: 𓄂—a noun meaning 'forepart (of an animal)', and which is always written ideographically with the front part of a lion and is transliterated *ḫꜣt*; and 𓃬𓍿𓏤—a noun meaning 'lion' and transliterated *mꜣi*. The writer uses the same visual wordplay in the following sentence, in the two words 𓈙—a noun meaning 'the hinder parts or hindquarters, which is always written with the hindquarters of a lion and is transliterated *pḫ*, and 𓃬𓍿𓏤—a noun meaning 'lion' and is transliterated *mꜣi*.

The writer here stresses his intended message visually and verbally, namely that the creature, which is required for this magical performance, should in some manner resemble the lion with its two special parts: the forepart and the hindquarters.

first word	second word	related images
𓄂	𓃬𓍿𓏤	𓄂 - 𓃬𓍿𓏤
𓈙	𓃬𓍿𓏤	𓈙 - 𓃬𓍿𓏤

Contrasted Words With the Same Determinative

The AE writer can use two contrasted words that employ the same soundless determinative, in order to stimulate the reader's mind about the sharp differences between the used words conceptually. In other words, this technique visualizes the existence and non-existence of the described object for the reader's eye:

𓈖𓏏𓅓�_𓐝𓏛𓇋𓅱𓅳𓏌𓏛𓈖𓏌𓐍𓅱𓏏

in wnm dp iw wšd.(t).w wšb.f

The one who eats, tastes; the one who has been questioned, answers. (Eloquent Peasant, Parkinson, 1991, pp. 33, l. 247)

Visual *jinās* is represented by using one shared ending determinative 𓀃, for four contrasted words in a stimulating way. The two words: 𓅳𓇋𓏌𓀃 *wšd*, 𓅳𓈖𓐍𓀃 *wšb*—meaning 'answering' and 'asking'. The determinative of the man putting his hand to his mouth 𓀃 refers here to the speaking activity. While the other two contrasted verbs 𓂝𓀃 *wnm*—�’𓏛𓀃 *dp*—meaning 'eating' and 'tasting' have used the same determinative 𓀃, to refer specifically to the activity of eating. This meaning is conveyed by the representation of the tongue in the verb *dpt* 𓄑 as an additional determinative representing the state of sensing the taste of the food, not the actual eating process. The writer here used *four* words sharing one image (man-with-hand-to-mouth 𓀃) as the sole dominant determinative to represent different actions related to the mouth. These visual contrasts are useful in comparing the actions of speaking and eating.

first word	second word	third word	fourth word	shared image
𓂝𓀃	�’𓏛𓀃	𓅳𓇋𓏌𓀃	𓅳𓈖𓐍𓀃	𓀃

Contrasted Words With Contrasting Determinatives

𓏭𓅓𓀀𓏥𓈗𓏏𓂝𓎼𓏤𓂝𓇋𓏏𓏤𓂋𓆰𓏥

sšmm.s ḥsw nb mi ḫt psft wꜣḏwt

She (the sky) warms everyone who is chilled like a fire that cooks raw things. (Eloquent Peasant, ibid., pp. 34, l. 277)

Visual *jinās* is represented by the employment of two contrasting determinatives that reflect two contrasted meanings: 𓏭𓅓𓀀𓏥 *sšmm*—a *sḏm.f* verb meaning 'to warm' and 𓈗𓏏𓂝𓎼𓏤 *ḥsw* a plural participle meaning 'people who feel cold'. Both words have two contrasting ending determinatives that illustrate the source of the described status: the first

word uses the fire determinative 🔥 to illustrate 'warmth' and 'heat', while the second word uses the boat's mast ⛵ which metaphorically represents the wind, but in this context, it represents a cold wind.

first word	second word	contrasted images
𓈖𓄿𓄿𓊮	�s𓂝⛵𓏥	🔥 - ⛵

The AE writer also visually reinforces the quality of being warmed by using a simile that employs two other words with a close relation to the fire determinative 🔥 of *sšmm*. The writer uses the words 𓐍𓏏 *ḫt*—meaning 'fire' and 𓊪𓋴𓆑𓏏 *psft*—meaning 'to cook', in order to stress the capability of the sky to warm everyone. The repetition of the fire determinative is pushing the idea of being warmed against the undesired condition of being cold in the visual context of the sentence.

first word	second word	third word	fourth word	visual sequence
𓈖𓄿𓄿𓊮	𓏿𓂝⛵𓏥	𓐍𓏏	𓊪𓋴𓆑𓏏	🔥 - ⛵ - 🔥 - 🔥

𓐍𓊪𓂋 𓄿𓅱 𓁷 𓅓 𓎛𓂝 𓄡𓏏 𓅓 𓇅𓄿 𓈖 𓈖𓏏𓏏 𓈖 𓇍𓏏 𓅓 𓄡𓂝𓅱 𓈖 𓈖𓏏𓏏 𓈖 𓐍𓊪𓂋𓏏

ḫpr ꜣw ḥr m ḥwꜥ ib m wꜣ n ntt n iit m ḫꜥw n ntt n ḫprt

The one who was happy (lit. with a joyful face) became like the one with grieved heart, do not scheme for something which did not come yet, do not rejoice for something which did not happen yet. (Eloquent Peasant, ibid., pp. 38, ll. 302–303)

Visual *jinās* is represented by employing two contrasting ending determinatives for the reversed *jinās* words: 𓎛𓂝𓅪 *ḥwꜥ*—meaning 'being sad', with a sparrow bird as a determinative that always represents negative meanings in the AE lexicon and 𓄡𓂝𓀁 *ḫꜥw*—being happy, with a man clapping or raising his hands as a determinative to express happiness. Both words use contrasted determinatives to illustrate the contradictory nature of the two emotional states better. The AE writing uses this small negative bird (the sparrow) to reflect the emotional status of being sad, this negative bird being metaphorically related to agricultural settings, where those small birds form a dangerous threat to the farmers when they devour their grains before they have a chance to grow and thus affect the crops produced. In the case of the word for happiness, AE writing uses a cheerful human figure.

first word	second word	contrasted images
𓎛𓂝𓅪	𓄡𓂝𓀁	𓅪 - 𓀁

Different Words With the Same Determinative

The visual features of AE writing allow its writers to employ two different words that are semantically related by using the same determinative, in order to illustrate metaphorical connotations that may exist between them. However, it should be stressed that the two words are not contrasted semantically.

in sḏrw mȝȝ rswt

    The sleeper is the one who can see the dreams. (Eloquent Peasant, Parkinson, 1991, pp. 33, l. 248)

Visual *jinās* is represented by using three successive, related determinatives to better illustrate the semantic context: *sḏrw*—a participle indicated by an image of a seated man, meaning 'the one who sleeps', 'lies down', 'goes to rest', or is 'inert', 'inactive', with a sleeping man over a bed as soundless determinative; then *mȝȝ*—a sḏm.f verb meaning 'to see', with an eye as a beginning hieroglyph; and finally *rswt*—a plural noun meaning 'dreams', with an eye as a determinative. The plural noun *rswt* is derived from the verb *rsw* which means 'being awake', and 'vigilance'. The writer thus considers the dreaming action during the sleeping process as a full, real, awakened state metaphorically. The shared eye determinative confirms this metaphorical similarity between the two contrasting actions. The link is related to the ability to see in both worlds, even with closed eyes during the sleeping process.

first word	second word	third word	related images

gm.n.i nb nṯrw iw m mḥy ṯȝw nḏm r ḫȝt.f šd.f sš-ḳd n imn nḫt imn mȝꜥ ḫrw

    I found the master of the gods coming like the north wind and the gentle breeze before him and he saved the draughtsman of the god Amun ꜥAmun-Nakhtꜣ the justified. (Neb Ra Hymn to Amun, Kitchen, 1980, pp. 653, ll. 8−9)

Visual *jinās* is represented by the visual relation between two words that have the same determinatives, artistically written beside each other: *mḥy*—a noun meaning 'north wind', 'storm that comes from the north', with a ship's mast as an ending soundless determinative, which metaphorically represents the strong wind and *ṯȝw*—a collective plural noun meaning 'wind', with a ship's mast as an ideogram

to metaphorically represent, in this context, the breeze or gentle wind. The creative writer highlights the ending determinative of the first word ⌐, by placing it beside another different word that has been written with a single symbol that stands for the whole word, which is the ship's mast 𓊬𓏲𓂝𓊬𓏼. The reading of its visual context may suggest that the writer aims, by using this visual correlation, at confirming two different notions related to the wind being super-fast but delicate and gentle as well, linking them metaphorically to his own praised god.

first word	second word	highlighted image
𓊬𓏲𓂝	𓊬𓏼	𓊬𓂝𓊬

Using an Unusual Determinative

The AE writer can change the usual ending soundless determinative to serve the described textual context innovatively. This writing technique shows how the visual memory of the reader's eye is important to decipher the intended message from the writer's side and how each different determinative can be pregnant with an additional layer semantically. The studied examples confirm that some AE words could bear two meanings at the same time as if one of them were winking at the other and the meaning overall lay in this wink. The same word, in a different textual context, can mean different things simultaneously. The AE writer can choose between the available determinatives that fit his textual context. Therefore, I would argue that changing the determinative is the dexterous performance of an innovative trick, by which one idea is presented when using the common word root, while another meaning is substituted by changing its usual determinative. This type of visual *jinās* gives more interesting answers about the AE writing practice and the authorial intention but also difficult questions related to the ancient reader's response to such visual instruments. It may raise questions about the impact of additional social factors on the conception of creating 'beautiful speech' within a specific discourse, such as social class, race, and even education level.

𓏼𓉐𓆟𓂋𓇋𓎼𓄿𓏏𓎛𓈖𓉐𓏤𓊖𓅱𓃀𓈖𓈙𓋴𓅱

ity n swḥꜣ n.f r ḳꜣ n pt ir **ḫryt** *Ꜥꜣt m tꜣ n šꜣsw*

    The sovereign of the one who gives praise for him to the height of heaven, who made great butchery in the land of Shasw. (Ramsess II Zigzag poem, Yoyotte, 1950, pl. VII, l. 6)

Visual *jinās* is represented by changing the common determinative of the word ���𓈖𓏏𓏥 *ḫryt*. The word is always written with a fallen cow with

its legs bound 🖾 that represents the slaughtering process of meat ani-
mals. The resourceful writer uses a fallen human determinative 𓀐 to
add a new visual meaning to the word. The new meaning confirms that
the slaughtering process is not anymore related to the animal world, but
rather to the enemy people of the Egyptian king in the land of Shasw.
This resourceful graphic change highlights the author's desire to get a
sense of double negative metaphors that convey the despising of the en-
emies of his praised king. On the one hand, the king approached his
enemy as the fearless butcher looms over the helpless animal intended
for slaughter; on the other hand, the enemy's resisting capability is not
better than the resistance of a bound fallen animal. In other words, the
enemy's status is nothing more than that of bound animals, ready to be
slaughtered, in the eyes of the mighty king.

This word 𓆷𓏏𓏲🖾 also corresponds with the similar word 𓆷𓀐 *ẖr*,
which is always used in military texts as a verb to mean falling down
or as a noun to mean the fallen enemy. The writer here depends on the
lexical memory of his readers to recall the common words in order to
decipher why he changes the word's usual determinative.

original word	used word	un/usual determinative
𓆷𓏏𓏲🖾	𓆷𓏏𓏲𓀐	𓀐 - 🖾

𓆓𓏏𓎡𓈙𓈖𓏲𓆊𓂝𓈖𓊨𓏏𓎡

*mk dmi.k **šnw** ḳȝ nst.k*

Look, your landing-stage is surrounded [by a crocodile] because of the truthfulness
of your tongue. (Eloquent Peasant, Parkinson, 1991, pp. 26, ll. 161–162)

Visual *jinās* is represented by the addition of an unusual determina-
tive, in order to add another semantic layer to the word. The word
𓈙𓈖𓆊 is derived from the verb 𓈙𓈖𓂝 *šnw*—meaning 'encircle' or the
verb 𓈙𓈖 *šn* meaning 'to circuit', with a determinative shaped like a cir-
cle or ring. The word 𓈙𓍢 *šnw*—meaning 'king's cartouche' is derived
from this verb as well. The writer uses the original root of the verb *šnw*,
but he added a crocodile's image 𓆊 as an ending soundless determina-
tive to create a new adjectival verb, meaning 'surrounded by crocodile'.
The crocodile determinative is visually and grammatically connected to
the previous word 𓂝𓅓𓏏𓊛 *dmi*—meaning 'harbour', 'quay'. The word ends
with a river channel as soundless determinative 𓈘. Both the crocodile
and the word *dmi* have been figuratively used here to indicate something
different from their literal meanings. The word *dmi* means here the an-
chorage of the afterlife paradise. Therefore, the determinative charac-
terizes the journey after death as a river expedition; while the crocodile
denotes evil doing or not telling the truth, to be more precise according
to the sentence context.

The visual message here is that the evildoer will find his evil-doing surrounds his harbour of paradise. The evil is metaphorically represented by the crocodile determinative. The writer here employs a horrible life experience that none of the receivers would wish to be part of in order to stress the importance of truthful speaking.[3] The writer uses the expression of the tongue's truthfulness as an implied critique from the robbed farmer to the government officials that he complains about. The visual context of these words has been used by the writer to reflect the power dynamic of this verse, which in turn encourages the reader to grasp what the farmer means by saying the opposite when using the crocodile's image.

first word	second word	related images

Conclusion

In alphabetical writings, the sound is partly dominent, as domenstrated in Arabic visual jinās. However, it becomes a secondary element in the case of AE visual jinās as it relies more on deciphering the implied message by the comprehension of 'seeing.' The individual ability of the AE writers, in using various eye-catching features to alert the mind's eye during the reading process, should not be overlooked or underestimated in both scripts (the hieroglyphic and its cursive version called hieratic). Using the understanding of our print culture, which is more related to rigid alphabetic constructions, should not overwhelm our critical minds in encoding the AE visual messages. AE writing reflects a different practice of using pictures like a true or metaphorical medium for literary communication. This picture has the potential to create new meanings or construct a visual argument for indigenous readers. (Rashwan, 2019, p. 154) Most of the scholarly focus goes then towards the verbal meaning and usual philological problems. Therefore, the visual aspects of the AE literary expression are still overlooked. The visual jinās is almost an ignored topic of investigation because scholars depend on the phonetic transliteration, which leaves out half of the artistic productivity of the AE writing system. Egyptologists became mechanically satisfied to replace the visual poetic form of the AE writing with deceptive transliterations.

---

3. The writer metaphorically uses a shared life memory of the AE culture to create a religious warning. Religions construct the afterlife punishments or rewards by extracting them from happy or painful life memories that humans experience during their earthly life. Literary exaggeration is always used to reinforce these religious ideas through creative literary devices.

References

Al-Bisāṭī, Mohammad [البساطي، محمود] (2018). التالد والطريف في فن جناس التصحيف
[The inherited and humorous in the art of visual jinas]. Ed. by Ashraf al-Iskandrany [أشرف الإسكندراني]. القاهرة :[Cairo] معهد المخطوطات العربية [The In-stitute of Arabic Manuscripts].

Darnell, John (2004). *The Enigmatic Netherworld Books of the Solar Osirian Unity: Cryptographic Compositions in the Tombs of Tutankhamun, Ramesses VI, and Ramesses IX*. Zurich: Orbis Biblicus Et Orientalis.

Diakonoff, Igor (1981). "Earliest Semites in Asia Agriculture and Animal Husbandry According to Linguistic Data (8th-4th Millennia BC)." In: *Altorientalische Forschungen* 8, pp. 23–74.

Ehret, Christopher (1995). *Reconstructing Proto-Afroasiatic (proto-Afrasian): Vowels, Tone, Consonants and Vocabulary*. Vol. 126. University of California Publications in Linguistics. Berkeley: University of California Press.

Greenberg, Joseph H. (1971). *Language, Culture and Communication: Essays*. Stanford: Stanford University Press.

Al-Gundy, Ali [الجندي، علي] (1954). فن الجناس: بلاغة – أدب – نقد *[The art of Jinās: Eloquence – Literature – Criticism]*. القاهرة :[Cairo] دار الفكر [House of Thought].

Hallof, Jochen (2011). "Esna." In: *UCLA Encyclopedia of Egyptology*. Los An-geles. URL: http://digital2.library.ucla.edu.

Hodge, Carleton (1983). "Afroasiatic: The Horizon and beyond." In: *The Jewish Quarterly Review* 74.2, pp. 137–158.

Kitchen, Kenneth A. (1980). *Ramesside Inscriptions, Historical and Biographical*. Vol. 3. Oxford: Blackwell.

Parkinson, Richard (1991). *The Tale of the Eloquent Peasant*. Oxford: Griffith Institute.

————— (1997). "The Text of 'Khakheperreseneb': New Readings of EA 5645, and an Unpublished Ostracon." In: *The Journal of Egyptian Archae-ology* 83, pp. 55–68.

Quirke, Stephen (2004). *Egyptian Literature 1800 BC: Questions and Readings*. London: Golden House.

Rashwan, Hany (2019). "Ancient Egyptian Image-Writing: Between the Unspoken and Visual Poetics." In: *Journal of the American Research Center in Egypt* 55, pp. 137–160.

————— (2020). "Comparative balāghah: Arabic and ancient Egyptian literary rhetoric through the lens of Post-Eurocentric Poetics." In: *The Routledge Handbook of Comparative World Rhetorics: Studies in the History, Application, and Teaching of Rhetoric Beyond Traditional Greco-Roman Contexts*. Ed. by Keith Lloyd. New York: Routledge, pp. 389–403.

Yoyotte, Jean (1950). "Les stèles de Ramsès II à Tanis." In: *Kêmi, Revue de philologie et d'archéologie égyptiennes et coptes* 11, pp. 47–62.

Mystic Messages—The Magic of Writing

Marc Wilhelm Küster

Abstract. This article studies how writing has been employed to pass subliminal messages not present in the spoken language, and how readers have searched for such secret messages.

The text illustrates how the intrinsic properties of both open and closed writing systems were used to create a magic of writing. It analyses examples from 1st millennium BC cuneiform acrostics to arcane readings of kanji in contemporary Japanese popular culture to show how writing is used as a non-linguistic system to encode and decipher messages.

For the first edition of *Grapholinguistics in the 21st Century*, the author has looked at open and closed writing systems. For him, open writing systems such as the cuneiform and Chinese writing systems are systems that have a core character repertoire, but can easily be extended (Küster, 2019). The character repertoires of closed writing systems such as the Greek or Latin alphabets, but also Hebrew abjads or the Hangul syllabary are essentially fixed.

This article further explores this the way open and closed writing systems operate and how in both cases writing can be a very different sign system from spoken language. This article looks at the magic of writing as an extreme case in which writing is imbued with meaning that its linguistic equivalent does not have. Already Küster (2006) studies acrostics, in which the order of the alphabet becomes a metaphor for completeness and ultimately perfection, and this remains one case covered. From here to number magic and Kabbalah it is but a small step. This glance at closed writing systems is then complemented by samples of contemporary Japanese culture. Three cases illustrate approaches to a magic of writing that work quite differently from the ones showcased for closed writing systems in the West.

Marc Wilhelm Küster   (iD) 0000-0002-2600-0717
3b, rue de Wormeldange
L-7390 Blaschette, Luxembourg
marc@budabe.eu

1. Magic

Probably already the earliest human societies sought ways of controlling the environment on which their survival depended— ensuring timely rain falls, good hunts, fertility, health and the clan's general prosperity. Magic, they believed, could grant them this power.

Often the performance of magic became the domain of specialists, be they called magicians, shamans, wise women, because it was imagined that its efficiency depended on precise adherence to a given set of often complicated ceremonies, invocations, incantations, and spells. Some speculate that already prehistoric cave paintings had been painted by shamans.[1] The potency of magic would often be reflected in its complexity; sloppiness in performing magic might result in the magician's death or worse and bring calamity to their societies, but if correctly executed great benefits could accrue:

> The magician does not doubt that the same causes will always produce the same effects, that the performance of the proper ceremony, accompanied by the appropriate spell, will inevitably be attended by the desired result. [...] If he claims a sovereignty over nature, it is a constitutional sovereignty rigorously limited in its scope and exercised in exact conformity with ancient usage.[2]

Magic existed for millennia, perhaps tens of millennia, before the invention of writing. It was therefore by necessity linked to spoken language and rituals. Its secrets were typically transmitted orally from teacher to student, with a lot of emphasis on precise repetition, though the cave paintings bear witness that even very early on practitioners seem to have wanted to persist aspects of their rituals and even thoughts and emotions.[3]

So it is maybe not surprising that over time writing was employed to record magic rituals, be they cuneiform tablets with incantations or (much later) grimoires and spell books.[4] However, these written representations of spoken magic that just reproduce oral spells are not the topic of this article, even though they might be an interesting study in an of itself, as many grimoires invented their own script-like devices to record magic.

---

1. Whitley (2009), who emphasises the "undeniable association between decorated panels and the deep, dark inner sanctums of the caves" (p. 28) and its links to shamans and shamanism.

2. Frazer (1890), Chapter IV. Magic and Religion—in this magic is truly the "bastard sister of science," in many ways its predecessor.

3. Whitley (2009) speaks of these paintings of "visionary images that illustrate the spirits and events of the supernatural world". If true, these images also would be among the earliest recorded instances of story telling.

4. See Davies (2009) for a history of spellbooks.

Instead, we look here at the mechanisms that were employed to transform writing itself into a vehicle for the performance of magic. We see how writing was used to pass subliminal messages not present in the spoken language and how readers searched for such secret messages, and at ways the intrinsic properties of writing were used for this—a magic of writing.

2.   The Origins of Magical Writing

As far as we know, cuneiform writing was quite prosaically invented as a mechanism for book keeping—a product of economic necessity to administer the increasingly complex commercial interactions in Sumerian city states. It was from its very beginning linked to ordered lists of terms as a key tool for memorising cuneiform characters and their hierarchical relation among each other and the concepts they represented.[5] At the beginning neither writing nor lexical lists seem have been linked to rites or magic, though by the Early Dynastic period we have first proof for incantations preserved on cuneiform tablets.[6]

2.1.   Acrostics

Going beyond recording incantations and spells that were originally transmitted orally, practitioners began to explore the internal logic of the semiotic system of first cuneiform and then alphabetic writing to create a magic of writing.

One of the early known ways of using a characteristic of writing for magical purposes was acrostic poetry where each line or alternatively each line in a stanza would begin with the same cuneiform character. While this might look like alliteration, it was not, as "[m]ost of the acrostics make use of this polyphony of sign values"[7] that marks many open writing systems. In other words, while the lines share the same initial character, they do not necessarily share the same phoneme.

The acrostic itself can transmit a hidden message:

---

5. Lexical lists are part of the earliest stage of script development in the Middle East, cf. Nissen (1981, p. 101) and Englund, Nissen, Damerow, and Baghdad (1993, p. 13).

6. http://etcsl.orinst.ox.ac.uk/edition2/literature.php— "Also first attested in the late Early Dynastic period are two particular types of non-utilitarian text that had a long history in Mesopotamia: incantations and royal inscriptions. The former employ various religious and rhetorical strategies, as well as mimetic ritual, to achieve instrumental ends such as curing illness."

7. Soll (1988, p. 307), here with reference to Babylonian acrostics.

Akkadian acrostics are message acrostics in which the initial syllables or signs of the horizontal acrostic lines spell out a message when read vertically. Seven complete or partial acrostic texts are known from Akkadian literature.[8]

In the case of the Babylonian *Theodicy*, written around 1000 BC, the message read "I, Saggil-kīnam-ubbib, the incantation priest, am adorant of the god and the king".[9] Other Akkadian texts would go even further and work with double acrostics with incantations encoded via both the initial and final syllables.[10] This sophisticated mechanism of hiding a message could only work in writing. It is the sheer difficulty of writing such acrostic poetry that in the mind of practitioners imbued it with its potency.

The Babylonian *Theodicy* is an acrostic in its written form—"[e]ach of the 11 lines of the stanzas start with the same sign, like Psalm 119. To achieve this [...] throughout the author allows himself the liberty of using a little the polyphony of the signs".[11] In other words, in contrast to alliteration the artifice cannot be separated from writing.

The advent of the alphabet with its intrinsic order would add an additional layer—acrostics whose first letters spelled out the order of the alphabet rather than a specific message. The alphabet itself is a "ready metaphor for totality"[12], a metaphor of completeness.

Alphabetic acrostics would become a staple also in wisdom literature, most prominently in a number of Hebrew psalms.[13]

2.2. *The Number of My Name*

One of the earliest known examples of number magic linked to a text—a name in this case—are foundation cylinders of the Assyrian king Sargon II (king from 722 to 705 BC) who had the wall around Dur-Sharrukin— "Sargon's fortress", today's Khorsabad in Iraq, about 15 km northeast of Mosul—built in a length that corresponded to "the number of his name".[14] Sargon II. proclaimed that "Vier Sar, drei Ner, 1 Soss, 3 kâné, 2 Ellen [Summa: 16280 Ellen]—so viel mein Name bedeutet—machte ich das Mass ihrer Mauer, und auf hohem Berggestein gründete ich fest ihr Fundament".[15]

---

8. See Brug (2010).

9. `http://www.etana.org/node/582`, cited after Lambert (1996, pp. 63–89).

10. Some such examples are the acrostics discussed in Sweet (1969).

11. See Lambert (1960, p. 66).

12. See Soll (1988, p. 317).

13. Cf. also W. G. E. Watson (1982) and W. G. Watson (1986).

14. At least four copies of this foundation cylinder are known, two each in the Louvre and in London, cf. Lyon (1883, p. XIV).

15. Lyon (1882, 11, line 65), also Lyon (1883, p. 39). The exact number is being discussed, with current literature going for 16,283 rather than 16,280 cubits, "De 16 283

The great city was entirely built in the decade preceding 706 BC. After the unexpected death of Sargon in battle in 705 BC, where even his body felt into the hands of his enemies (an extremely bad omen in Assyria), Dur-Sharrukin was abandoned and the capital was shifted 20 km south to Nineveh.

While there is no consensus yet on how this calculation was made and which exact numerical values should be applied for the cuneiform characters in the case of Sargon II's name, we can safely assume that the Assyrian king expected this message on Dur-Sharrukin foundation cylinders to magically strengthen the wall of his newly founded capital. Presumably the mystic equivalence of the royal name and his capital's wall was conceived to link and reinforce both. It may well be that, when Sargon II ignominiously died in battle shortly afterwards, it was this same equivalence that helped to doom his short-lived capital.

Sargon II could build on a tradition of cuneifom signs that were mapped to numbers. As Pearce (1996) elaborates in his article on *number-syllabary texts* in that late cuneiform period, the numeric value was one linked to the cuneiform signs themselves, not to their typically polyvalent readings.[16] What is more, these mappings were not secret knowledge, not *Geheimwissen* (cf. p. 461), but rather accepted knowledge that became soon loaded with theological overtones—"by the first millennium, numerals frequently represented divine names."[17]

To map numbers and characters is not obvious in an open writing systems, whereas it is a perfectly natural feature of a closed writing system with a clear order between its letters. On a purely speculative note, given the very late attestation of number-syllabary tests in the history of the cuneiform script at a time where the Phoenician and Aramaic scripts were long in current use, it could very well be inspired by the numerology of the early Semitic scripts. It would then naturally have fit in a existing tradition of acrostics and other forms of sign magic.

3.   Kabbalah

Just like alphabetic acrostics, the Kabbalah and in particular its use of gematria to search for hidden messages in the Torah derives from the double use of Greek and Hebrew letters as number signs, which in turn reflects the order of those letters in their writing systems.

---

grandes coudées, le nombre de mon nom, je fis le circuit de sa muraille", cf. Contenau, 1940, p. 162. / "I made the circumference (lit., measure) of its (the city's) wall 16,283 cubits, (corresponding to) my name (*nibīt šumīya*), and established the foundation platform upon the bedrock of the high mountain" Frahm (2005, p. 48).

16. The texts "demonstrate that the scribes intended the numerals to represent only the sign form and not the possible syllabic readings of the sign" (Pearce, 1996, p. 460).

17. See Pearce (ibid., p. 461).

As we have seen above, the idea of finding hidden numbers and meanings in sequences of letters seems to have a long tradition in the Middle East, with Sargon II's magic just being a case in point. Number magic cut two ways—practitioners used it to imbue texts with magic by consciously encoding into them further layers of meaning, but also tried to decypher such layers from texts that certainly were never meant to have them.

These strategies were known in Israel in the Hellenistic period, but are likely much older—as mentioned above, it might have been itself the inspiration for applying number magic in the late cuneiform period.[18] At a time where the Aramaic writing system was fully established, some Aramaic texts were even "retrofitted" into cuneiform to strengthen this link.[19] Similarly, Lieberman (1987, p. 167) traces Rabbinic gematria squarely back to Mesopotamian practices.

3.1. The *Zohar*

While some apologists today claim that the Kabbalah go back to Talmudic, if not Mosaic or even Adamic times,[20] it probably came about only in the 12th century, though building on a much older tradition of Jewish mysticism. Its foundational work, the *Zohar*, was written—or, as some practitioners would have it, rediscovered—in 13th century Spain.

Among many other things the *Zohar* developed an outright philosophy of the role that the Hebrew letters had in the creation of the world, starting out in the prologue where "When He desired to create the world, all the letters came before Him in sequence from last to first"[21] with Bet ב being the letter chosen to create the world "because the letter Bet is the first letter of the word blessing (HEB. BRACHAH)"[22]. Here, the Torah starts with Bet. Aleph א is compensated by the honour to "be the first [...] of all the letters [...] all calculations and actions of the people shall commence with you. Therefore, all unity shall be expressed by the letter Aleph!"[23]

---

18. For an interesting, but explicitly speculative study on occurrences of 52—the numerical value associated with the name of god—already in the Masoretic version of the Torah cf. Knohl (2012).

19. Cf. Gordon (1937) for an example of a late cuneiform incantation tablet written in Aramaic. While the tablet itself was written only in Hellenistic times, the incantation itself may be older, since "[i]t is well known that the efficacy of an incantation is often believed to be in direct proportion to its antiquity," p. 105.

20. Cf., e.g., Kurzweil (2007, 35ff).

21. Zohar, Prologue, verse 23, cited after https://www.zohar.com/zohar/Prologue/chapters/6.

22. Loc. cit., verse 37, capitalisation in the original.

23. Loc. cit., verse 38.

It could be the topic of another article to look more in detail at the grafematics of the *Zohar* and of the Kabbalah as a whole. Here, let it be sufficient to state that this philosophy does not stop at the letters itself. The forms of the letters—in this case the stroke of Aleph א—is also loaded with meaning:

> Come and see: The first subject of the Torah we give to children is the Alphabet. […] Even supernal angels and the most sublime cannot comprehend it, as these matters are the mysteries of the Holy Name. There are 14,050,000 worlds dependent upon the stroke of the Aleph א, MEANING THE STROKE OF THE UPPER YUD OF THE ALEPH, and 72 holy names are engraved in the impressed letters in them. The high and low beings; heaven, earth and the seat of glory of the King—are hanging from one side to the other side, MEANING FROM THE UPPER STROKE TO THE LOWER STROKE of the expansion of the Aleph […][24]

Given the central role of Hebrew letters and writing in general, it is not surprising that the Kabbalah sought to extract layers of meaning from the Torah that are linked to the written word itself—"Kabbalists believe that God created the world through a combination of Hebrew letters"[25], as a contemporary popular introduction to the Kabbalah puts it. The classical layer of interpretation went from the "literal meaning, the homiletic meaning, the hints that the text implies, and the secret, mystical meaning"[26]. For this latter gematria is a common, though by no means not the only approach to extract these type of mystical meanings through the contemplation of the—in a literal sense—written word.

This love for letters also inspired a number of non-Jewish practitioners of the Kabbalah—while the Kabbalah originates in Judaism, it had adepts also in other religions. The Christian scholar Ramon Lull was one of the most prominent of them, searching in the spirit of the ecstatic Kabbalah, "combinations of letters which constitute the Divine name"[27], but by far not the only one. Even today gematria is still being used in some esoteric circles in the hope of extracting deep hidden messages from written texts.

3.2. Kabbalah and True Names

The Middle Eastern tradition met with the Socratic search for *true names* and their numeric identities. In Greece Plato's ideas would give additional impetus, as in *Cratylos* where Plato makes Socrates say:

---

24. Zohar, Acharei Mot, verse 303, cited after `https://www.zohar.com/zohar/Acharei\%20Mot/chapters/50`. Capitalisation in the original.

25. See Kurzweil (ibid., p. 130).

26. See Kurzweil (ibid., p. 218).

27. See Idel (1988, p. 171).

On this basis, then, you will judge the law-giver, whether he be here or in a foreign land, so long as he gives to each thing the proper form of the name, in whatsoever syllables, to be no worse lawgiver, whether here or anywhere else, will you not? (Plat. Crat. 390a)

This hunt for the *proper form*, the *true name*[28] of a thing became common in the Hellenistic and post-Hellenistic world and well beyond. It influences strands of the Kabbalah, which elevated the concept of finding hidden meanings into a philosophy. In particular gematria became a major tool. The ultimate meaning of a text would not be the one indicated by the language that the writing ostensibly represents, but rather a mystical message transmitted via numbers encoded exclusively in the written word.

4.   Magic in Open Writing Systems

Gematria and alphabetic acrostics work best for closed writing systems. They depend on a mapping between letters and numbers, typically via the supposedly eternal alphabetic order of letters in the writing system.[29]

Though open writing systems also partially adopted these techniques, they had to find also other techniques to elicit hidden meanings from a written text. Contrary to Sumerian writing, the oldest examples of notably Chinese characters are actually directly linked to magic, more specifically pyromancy, divination by fire. Questions would be carved on bones in what is the oldest surviving form of Chinese writing from around 1200 BC. The diviner—often the king in person—would then heat the bone and interpret the resulting cracks to elicit an answer to the question. Sometimes this interpretation would be noted on the bone to compare it with the actual event later.

In Chinese culture this link between writing and ritual would never be fully broken, a trait that also of the Japanese writing system inherited. In Japan ema—wooden plaques—became a standard way to transmit wishes to the gods. The wish is inscribed on the plaque and hung at the shrine in question before being burned. Its written message is thereby posted to the god to whom the shrine is dedicated. This practice is still very much alive—some shrines figure prominently also in popular culture and some ema even figure characters from Japanese popular culture.[30]

In this article I will look at three contemporary examples from popular culture that exploit properties of Japanese writing:

28. The latter a concept explored, e.g., in Le Guin's Earthsea cycle, Le Guin (2012).
29. See Küster (2006, 181ff).
30. As described in Reader (1991), one such example is the real-life Shirakawa Hachiman Shrine that became a target of pilgrimage for fans of the *Higurashi—When They Cry* visual novel and anime franchise. The fans celebrate this shrine as the ficti-

4.1. Magic in Polyphony

Magic in Hayao Miyazaki's animated masterpiece *Spirited away* features the ability of the witch Yubaba to enslave the humans that request to work for the spirit world bathhouse which she manages. She takes away their liberty and their identity of her applicants by re-baptizing them. However, this power seems constrained; she can only operate on a written contract and shorten her victim's written name, allowing her workers a tenuous link to their previous existence.

This specific constraint is lost in the English translation, where the name of the protagonist, Ogino Chihiro, inexplicably morphs into Sen. The logic behind this it is much more visible in the original Japanese title, 千と千尋の神隠し, *Sen and Chihiro's Spiriting Away*, and it is immortalized in the scene where Chihiro is forced to sign away her name and with it her identity. Yubaba makes all parts of Chihiro's signature disappear except for the single character 千, the first kanji in Chihiro's first name 千尋:

FIGURE 1. Miyazaki (2001), position 38:41

Like most Japanese kanji, 千 is the polyphonic grapheme which has two main syllabic readings, *chi* (the Japanese *kun* reading), but also *sen* (the "Chinese" *on* reading of the character).

Yubaba's magic thus relies on her ability to operate through writing and on *written* words. It exploits the polyphony of Japanese characters

---

tious Furude Shrine in the equally fictitious village of Hinamizawa, addressing ema to some of the *Higurashi* characters. Andrews (2014) studies other examples.

to manipulate associations between identity, the written and the spoken word in ways that transcend spoken language.

Furthermore, by choosing the On-reading for Chihiros new name, Yubaba seems to remove Chihiro from her true Japanese roots. In addition, 千 also stands for 1,000, an implicit statement of capitalistic values usurping Chihiro's true personality. She has to start on a long inquiry[31] to reestablish her true identity and values.

4.2.   Radical Magic

The *Monogatari* series of Japanese popular novels, authored by Nishio Isin, often situates itself with reference to a Shintoist pantheon and Shintoist magic. This magic very often turns around more esoteric interpretations of writing in general and kanji in particular. Even the animated rendering of the novels regularly showcases tablets of written texts to underline links that the spoken language cannot show.

One such story centres around helping a desperate girl, Hachikuji Mayoi, trying to find her mother. She claims to be a lost snail:

FIGURE 2. Itamura (2009), position 22:44. In the series written passages like this one regularly complement the story.

---

31. The second kanji 尋 of Chihiro's first name is also the stem of the verb 尋ねる, to inquire, though now with the alternative reading *tazu* rather than *hiro*—another association that only works in the written language.

蝸牛 is the spelling that the story adopts for *snail*; it is a less common out of a number of possible forms of writing the underlying phonemes.[32] However, it is only on the choice of this particular spelling on which the magic works:

- The first of the two kanji, 蝸, contains the radical 咼 for *evil* and *dishonest*, associating the girl with negative forces
- The second of the kanji, 牛, symbolises *cow*, situating the story in a family of myths of entities who lure innocent wanderers astray

These same plays on specifically the written language are equally prevalent in the source material (Nishio, 2006).

The magic is in the letters themselves—by interpreting the underlying kanji characters in a arcane way, the author manages to associate the innocent and at first glance innocuous concept of a lost snail with other, more established ideas in myth.

4.3.   The Servant's Three New Names

In Adachi, 2014, the Shinto god Noragami binds a new servant ("shinki") to his services. He literally inscribes a "property mark" in the form of a kanji on his servant and joins this with three different readings—his everyday name Yuki, his functional name, Setsu, and a third, true name that is not voiced. It is ultimately this third, secret reading of the servant that only he as a god can sense and that ultimately gives him power over his servant.

While coming from a very different cultural background, the idea of this third, hidden name bears strong similarity to the Platonic concept of a person's or object's true name. Finding this true name is envisioned to confer power over the named person or object—a vision obviously shared between Shintoism, much of the neo-platonic school, and the Kabbalah.

5.   Summary

Using a number of examples this article demonstrate how the intrinsic properties of writing systems can carry mystical messages in the mind of practitioners. These messages depend only an esoteric interpretation of the written language that hides or reveals these messages.

The concrete methods of doing so depend on the logic of the writing system in question. In our selected examples we have encountered:

---

32. Cf.   https://www.linguee.com/english-japanese/translation/snail.html, last consulted on 2020-01-19.

FIGURE 3. Adachi (2014, p. 177)

- acrostic poetry that imbues its text with messages hidden in its first and sometimes also last characters;
- number magic through numeric values associated with characters that are supposed to encode a text's deeper, secret truths, which the practitioner hopes to exploit or to which he hopes to gain access;
- links created between concepts by exploiting the internal structure of characters;
- polyphony that allows multiple readings of characters, creating again associations and messages that are pure artefacts of writing.

These techniques may not be the most prevalent use of writing. However, they demonstrate how writing systems can (partially) emancipate themselves from their underlying spoken languages and find new, non-linguistic ways to transmit meaning. For closed writing systems these mechanisms mainly exploit the writing system's internal structures, whereas open writing systems rely more on arcane features of the characters themselves. However, in both cases they succeed in creating a new system—a magic of writing.

References

Adachi, Toka (2014). *Noragami: stray god*. New York: Kodansha.

Andrews, Dale K. (2014). "Genesis at the Shrine: The Votive Art of an Anime Pilgrimage." In: *Mechademia* 9, pp. 217–233.

Brug, John (2010). "Near Eastern Acrostics And Biblical Acrostics Biblical Acrostics And Their Relationship To Other Ancient Near Eastern Acrostics." In: URL: https://essays.wls.wels.net/bitstream/handle/123456789/856/BrugAcrostics.pdf.

Contenau, G. (1940). "Notes d'iconographie religieuse assyrienne." In: *Revue d'assyriologie et d'archéologie orientale* 37.4, pp. 154–170. URL: http://www.jstor.org/stable/23294662.

Davies, Owen (2009). *Grimoires: a history of magic books*. Oxford New York: Oxford University Press.

Englund, R. K et al. (1993). *Die lexikalischen Listen der archaischen Texte aus Uruk:* Archaische Texte aus Uruk. Gebr. Mann. URL: http://books.google.com/books?id=q6aFQgAACAAJ.

Frahm, Eckart (2005). "44 Observations on the Name and Age of Sargon II and on Some Patterns of Assyrian Royal Onomastics." In: *Nouvelles Assyriologiques Brèves et Utilitaires* 2.

Frazer, James George (1890). *The Golden Bough: a study of magic and religion*. London: Macmillan and Co.

Gordon, Cyrus H. (1937). "The Aramaic Incantation in Cuneiform." In: *Archiv für Orientforschung* 12, pp. 105–117.

Idel, Moshe (1988). "Ramon Lull and Ecstatic Kabbalah: A Preliminary Observation." In: *Journal of the Warburg and Courtauld Institutes* 51, pp. 170–174.

Itamura, Tomoyuki (2009). まよいマイマイ其ノ壹 *[Mayoi Snail, Part One]*.

Knohl, Israel (2012). "Sacred Architecture: The Numerical Dimensions of Biblical Poems." In: *Vetus Testamentum* 62.2, pp. 189–197. URL: http://www.jstor.org/stable/41583730.

Kurzweil, Arthur (2007). *Kabbalah for dummies*. Hoboken, NJ: Wiley Pub.

Küster, Marc Wilhelm (2006). *Geordnetes Weltbild*. Tübingen: Niemeyer.

———— (2019). "Open and Closed Writing Systems. Some Reflections." In: *Proceedings of Graphemics in the 21st Century, Brest 2018*. Ed. by Yannis Haralambous. Brest: Fluxus Editions, pp. 17–26.

Lambert, Wilfred G. (1960). *Babylonian Wisdom Literature*. 1st ed. Oxford: Clarendon Press.

———— (1996). *Babylonian Wisdom Literature*. 2nd ed. Winona Lake: Eisenbrauns.

Le Guin, Ursula (2012). *The Earthsea quartet*. London: Viking.

Lieberman, Stephen J. (1987). "A Mesopotamian Background for the So-Called Aggadic 'Measures' of Biblical Hermeneutics?" In: *Hebrew Union College Annual* 58, pp. 157–225.

Lyon, David Gordon (1882). *Die Cylinder-Inschrift Sargons II in transscribirtem assyrischem Grundtext mit Übersetzung und Commentar*. Hundertstund & Pries.

———— (1883). *Keilschrifttexte Sargon's Königs von Assyrien (722-705 v. CHR.), nach den Originalen neu hrsg., umschrieben, übersetzt und erklärt*. Hinrichs.

Miyazaki, Hayao (2001). 千と千尋の神隠し *[Spirited Away]*. 株式会社スタジオジブリ [Studio Ghibli].

Nishio, Ishin (2006). *Monster Tale*. Vol. 1. Bakemonogatari. Tokyo: Kodansha.

Nissen, Hans J. (1981). "Bemerkungen zur Listenliteratur Vorderasiens im 3. Jahrtausend (gesehen von den Archaischen Texten von Uruk)." In: *La lingua di Ebla. Atti del convegno internazionale (Napoli, 21-23 aprile 1980)*. Ed. by Luigi Cagni. Napoli, pp. 99–108.

Pearce, Laurie E. (1996). "The Number-Syllabary Texts." In: *Journal of the American Oriental Society* 116.3, pp. 453–474.

Reader, Ian (1991). "Letters to the Gods: The Form and Meaning of Ema." In: *Japanese Journal of Religious Studies* 18.1, pp. 24–50.

Soll, William Michael (1988). "Babylonian and Biblical Acrostics." In: *Biblica* 69.3, pp. 305–323.

Sweet, R. F. G. (1969). "A pair of Double Acrostics in Akkadian." In: *Orientalia* 38.3, pp. 459–460.

Watson, Wilfred G. E. (1982). "Trends in the development of classical Hebrew poetry. A comparative study." In: *Ugarit-Forschungen* 14, pp. 265–277.

———— (1986). *Classical Hebrew poetry—a guide to its techniques*. Journal for the study of the Old Testament. Sheffield: JSOT Press.

Whitley, David (2009). *Cave paintings and the human spirit: the origin of creativity and belief*. Amherst, NY: Prometheus Books.

Index

Q

R

U