

GRAPHOLINGUISTICS AND ITS APPLICATIONS

# Grapholinguistics in the 21st Century—2022

/gʁafematik/  
*Proceedings*

June 8-10, 2022, Palaiseau  
Yannis Haralambous (Ed.)

## Part I

Fluxus Editions



# Grapholinguistics and Its Applications 9



# Series Editor

Yannis Haralambous, *IMT Atlantique & CNRS Lab-STICC, France*

# Series Editorial Committee

Gabriel Altmann†, *formerly Ruhr-Universität Bochum, Germany*

Jacques André, *formerly IRISA, Rennes, France*

Vlad Atanasiu, *Université de Fribourg, Switzerland*

Nicolas Ballier, *Université de Paris, France*

Kristian Berg, *Universität Oldenburg, Germany*

Chuck Bigelow, *Rochester Institute of Technology, USA*

Stephen Chrisomalis, *Wayne State University, USA*

Florian Coulmas, *Universität Duisburg, Germany*

Joseph Dichy, *Université Lumière Lyon 2 & CNRS, Lyon, France*

Christa Dürscheid, *Universität Zürich, Switzerland*

Martin Dürst, *Aoyama Gakuin University, Japan*

Keisuke Honda, *Imperial College and University of Oxford, UK*

Shu-Kai Hsieh, *National Taiwan University, Taiwan*

Terry Joyce, *Tama University, Japan*

George A. Kiraz, *Institute for Advanced Study, Princeton, USA*

Mark Wilhelm Küster, *Office des publications of the European Union, Luxembourg*

Gerry Leonidas, *University of Reading, UK*

Dimitrios Meletis, *Universität Zürich, Switzerland*

Kamal Mansour, *Monotype, USA*

Klimis Mastoridis, *University of Nicosia, Cyprus*

Tom Mullaney, *Stanford University, USA*

Martin Neef, *Technische Universität Braunschweig, Germany*

J.R. Osborn, *Georgetown University, USA*

Cornelia Schindelin, *Johannes Gutenberg-Universität Mainz, Germany*

Virach Sornlertlamvanich, *SICCT, Thammasat University, Thailand*

Emmanuel Souchier, *Université de la Sorbonne, Paris*

Jürgen Spitzmüller, *Universität Wien, Austria*

Richard Sproat, *Google, USA*

Susanne Wehde, *MRC Managing Research GmbH, Germany*



Yannis Haralambous (Ed.)

# Grapholinguistics in the 21st Century

/gɾafematik/

June 8–10, 2022 Palaiseau, France

Proceedings

Part I

**Fluxus Editions**



Yannis Haralambous (Ed.). 2024. *Grapholinguistics in the 21st Century. June 8–10, 2022. Proceedings* (Grapholinguistics and Its Applications, Vol. 9). Brest: Fluxus Editions.

This title can be downloaded at:

<http://fluxus-editions.fr/gla9.php>

© 2024, The respective authors

Published under the Creative Commons Attribution 4.0 License

(CC BY 4.0): <http://creativecommons.org/licenses/by/4.0/>

ISBN: 978-2-487055-04-9

e-ISBN: 978-2-487055-05-6

ISSN: 2681-8566

e-ISSN: 2534-5192

Cover illustration: Part of “Ensemble de quatre dessins,” 1976, ink on paper, by Marcelle Cahn (Strasbourg 1895 – Neuilly-sur-Seine 1981), Strasbourg Modern and Contemporary Art Museum. (To discover Marcelle Cahn, watch the beautiful documentary <https://www.rts.ch/play/tv/-/video/-?urn=urn:rts:video:12984455> broadcasted on May 19th, 1976 from RTS as part of the *Clés du regard* program.)

Marcelle Cahn: *Composition non figurative* (pour l’œuvre 55.980.2.44). Musée d’Art Moderne et Contemporain de Strasbourg, Cabinet d’Art Graphique

Photo Musées de Strasbourg, M. Bertola

Droits réservés.

Cover design and typesetting: Atelier Fluxus Virus

Main fonts: William Pro by Typotheque Type Foundry, Computer

Modern Typewriter by Donald E. Knuth, Source Han Serif

by Adobe Systems, Amiri by Khaled Hosny

Typesetting tools: X<sub>Y</sub>L<sup>A</sup>T<sub>E</sub>X, biblatex+biber (authoryear-icomp style),

[xindex, titlecaseconverter.com](http://xindex.titlecaseconverter.com)

Fluxus Editions

38 rue Émile Zola

29200 Brest, France

[www.fluxus-editions.fr](http://www.fluxus-editions.fr)

Dépôt légal: avril 2025

λγκε



# Table of Contents

|  |    |
|--|----|
| <i>List of Participants at the Grapholinguistics in the 21st Century<br/>2022 Conference</i> . . . . . | ix |
|--|----|

## PART I

|   |     |
|---|-----|
| DIMITRIOS MELETIS. – What’s in a name? Trends and challenges<br>in naming the study of writing . . . . .  | 1   |
| AMALIA E. GNANADESIKAN. – Amodal Morphology. Applications<br>to Brahmic Scripts and Canadian Aboriginal Syllabics . . . . .                                       | 45  |
| PAOLO COLUZZI. – The ideology of “monographism” and the ad-<br>vantages of digraphia . . . . .  | 67  |
| LIUDMILA L. FEDOROVA & ANTONIO PERRI. – Emblematic tech-<br>niques as textual strategies in non-linear and linear scripts . .                                     | 75  |
| MARY C. DYSON. – Perceptual Disfluency Through Hard-to-Read<br>Fonts. Is There a Satisfactory Explanation? . . . . .  | 101 |
| CHRISTINE KETTANEH. – Asemic Writing, Homebound . . . . .   | 111 |
| TEREZA SLAMĚNÍKOVÁ. – Sinograms on Commercial Signs. A Case<br>Study of Chinese Restaurants in Prague . . . . .   | 135 |
| CHENCHEN SONG. – Sentence-Final Particle vs. Sentence-Final<br>Emoji. The Syntax-Pragmatics Interface in the Era of Computer-<br>Mediated Communication . . . . . | 157 |
| DANIEL HARBOUR. – The Rosetta Stone Squandered: Decipher-<br>ment’s Twelve-Year Gap and the Fate of J.D. Åkerblad . . . . .                                       | 193 |
| SVEVA ELTI DI RODEANO. – From Clay Tablet to Digital Tablet.<br>The Diamesic Variation of Writing . . . . .   | 219 |



|  |     |
|--|-----|
| KRISTIAN PASKOJEVIĆ. – The application of grapholinguistics in palaeography. A case study: Croatian Glagolitic and Cyrillic palaeography . . . . . | 237 |
| KATHARINA TYRAN. – Reinterpreting the Semiotics of Glagolitic  | 253 |
| TIMO HOMBURG & THIERRY DECLERCK. – Towards the integration of cuneiform in the OntoLex-Lemon framework . . . . .                                   | 265 |
| JANUSZ S. BIEN. – 16th century Latin printed brevigraphs in Unicode—a computer resource . . . . .  | 299 |
| RAIOMOND DOCTOR, ALEXANDER GUTKIN, CIBU JOHNY, BRIAN ROARK & RICHARD SPROAT. – Graphemic Normalization of the Perso-Arabic Script . . . . .        | 315 |
| JOSEPH DICHY. – Semitic Writings and Short Vowels: Alternative Hypotheses in a Renewed View of the Analytics of Writing . .                        | 377 |
| DANA AWAD. – Reasons for Re-Paragraphing in the Translation Process . . . . .  | 391 |

## PART II

|   |     |
|---|-----|
| MICHAL SHOMER. – Introducing Multi-Gender Hebrew . . . . .  | 399 |
| ARVIND IYENGAR. – The akshara as a graphematic unit . . . . .   | 419 |
| RACHEL GARTON, MERRION DALE, L. SOMI ROY & PRAFULLA BASUMATARY. – Endangered Languages in the Digital Public Sphere. A case study of the writing systems of Boro and Manipuri . . . | 437 |
| GORDON BERTHIN. – Qualitative and Quantitative Validation of <i>Rongorongo</i> Glyph Strings on Easter Island Artefacts . . . . .   | 471 |
| TOMI S. MELKA & ROBERT M. SCHOCH. – The Intersection between Art, Non-Linguistic Symbol Systems, and Writing. The Case of the Wari, Tiwanaku, and Inka Iconographies . . . . .      | 501 |
| ZOFIA JANINA BORYSIEWICZ. – Life of Chaim . . . . .   | 611 |
| KAMAL MANSOUR. – The Sorcerer's Brew. The unexpected Results of Typographic Innovation . . . . .  | 625 |
| MARC WILHELM KÜSTER. – Fantastic Letters—Writing in a Fictional World . . . . .   | 639 |
| HELEN MAGOWAN. – De-aestheticizing the Artist's Brush. Calligraphy Manuals and the Pragmatics of Calligraphic Writing . .   | 653 |

---

|   |     |
|---|-----|
| PIERRE MAGISTRY & YOANN GOUDIN. – Semanticity in the Chinese Graphic System. Modeling and Assessing its consistency . . .   | 671 |
| ELVIN MENG. – Zheng Qiao’s Grammatology . . . . .   | 689 |
| CORNELIA SCHINDELIN. – The Chinese Script as a Self-Regulating System. Applying Köhler’s Basic Model of Synergetic Linguistics to Simplified Chinese Characters . . . . .   | 739 |
| HANA JEE, MONICA TAMARIZ & RICHARD SCHILLCOCK. – Does Korean grapho-phonemic systematicity enhance spontaneous learning? . . . . .  | 771 |
| ADRIEN CONTESSE, MORGANE RÉBULARD, CHLOÉ THOMAS, CLAUDIA S.BIANCHINI, CLAIRE DANET, LÉA CHEVREFILS, PATRICK DOAN. – Concevoir une fonte pour la transcription des mouth actions en langue des signes. Le système typographique Typannot . . . . . | 781 |
| <i>Index</i> . . . . .  | α’  |





# List of Participants at the *Grapholinguistics in the 21st Century 2022* Conference

Awad, Dana  
Baize-Varin, Marie  
Berthin, Flora  
Berthin, Gordon  
Bianchini, Claudia  
Bień, Janusz S.  
Borysiewicz, Zofia Janina  
Bouilleaud, Nicolas  
Coluzzi, Paolo  
Contesse, Adrien  
Cornelia, Schindelin  
Dale, Merrion  
Danet, Claire  
Dichy, Joseph  
Dunlavey, Nicholas  
Dürst, Martin  
Dyson, Mary  
Elti di Rodeano, Sveva  
Farrando, Pere  
Fedorova, Liudmila  
Fetnaci, Nawal  
Filhol, Michael  
Furner, Jonathan  
Garton, Rachel  
Gnanadesikan, Amalia  
Gorman, Kyle  
Goudin, Yoann  
Gutkin, Alexander  
Habibifar, Elnaz  
Haralambous, Yannis  
Harbour, Daniel  
Homburg, Timo  
Honda, Keisuke  
Humberstone, Katy  
Issele, Joanna  
Iyengar, Arvind  
Jarosch, Julian  
Jee, Hana

Joyce, Terry  
Karakılçık, Pınar  
Kerschhofer-Puhalo, Nadja  
Kettaneh, Christine  
Kučera, Jan  
Küster, Marc Wilhelm  
Magistry, Pierre  
Magowan, Helen  
Mansour, Kamal  
McCay, Kelly  
Melka, Tomi  
Meletis, Dimitrios  
Meng, Elvin  
Novel, Grégoire  
Osborn, J.R.  
Osterkamp, Sven  
Paskojević, Kristian  
Perri, Antonio  
Rébulard, Morgane  
Salomon, Corinna  
Schindelin, Cornelia  
Schoch, Robert  
Schreiber, Gordian  
Shomer, Michal  
Simpson, Logan  
Skandalis, Maximos  
Slaměniková, Tereza  
Song, Chenchen  
Spitzmüller, Jürgen  
Sproat, Richard  
Sugimori, Noriko  
Taha, Haitham  
Thomas, Chloé  
Tian, Tian  
Tyran, Katharina  
von Ascheberg, Thomas  
Wąsowicz-Peinado, Aleksandra  
Xu, Duoduo





# What's in a Name?

## Trends and Challenges

### in Naming the Study of Writing

Dimitrios Meletis

*Abstract.* The name of a scientific discipline is closely tied to the discipline's definition and (self-)conception. This renders naming processes highly significant as they involve intricate negotiations of and ultimately decisions concerning, among many other aspects, the boundaries of the newly designated discipline and research traditions that the chosen label may be associated with. In the little-researched history of the study of writing, scholars have proposed several names at different times and in diverse contexts. In this historiographic paper, nine are discussed: *grammatology*, *graphonomy*, *graphology*, *graphem(at)ics*, *orthography*, *writing systems research*, *grapholinguistics*, *script(ur)ology*, and *philography*. The 'baptism stories' behind these designations are characterized by common trends and challenges arising from the goal of coining a semantically transparent and unambiguous term that fits the study of writing and is more or less inclusive of the multiple disciplines and perspectives that wish to participate in it. Given that no name has been widely adopted and processes of disciplinary demarcation are still ongoing, this paper aims to systematically shed light on this important if somewhat chaotic part of the history of the study of writing to raise awareness and ultimately inform future efforts in (further) establishing it.

Names matter. They are not only labels or reference terms for historical accounts, but strategic tools.

---

De Chadarevian (2002, p. 206)

Nomenclatural questions [...] should, in any case, detain us only in idle moments.

---

Watt (1994a, p. xii)

## 1. The Goal

"What's in a name? That which we call a rose by any other name would smell just as sweet."—William Shakespeare's famous line from *Romeo and Juliet* implies that the naming of things is arbitrary, that their intrinsic

---

Dimitrios Meletis  0000-0002-8889-6459

Department of Linguistics, University of Vienna, Sensengasse 3A, 1090 Vienna, Austria. E-mail: dimitrios.meletis@outlook.com

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 1–44. <https://doi.org/10.36824/2022-graf-mele>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

qualities are not captured by labels.<sup>1</sup> Given the arbitrariness of linguistic symbols, most linguists would certainly agree with this assessment with respect to ‘common words’ used in everyday language. The story is arguably different for technical terms, to which scholars regardless of their discipline commonly ascribe great relevance—especially when the terms are meant to label entire branches of study. One reason for this is that such designations are products of conscious and complex naming processes, which themselves become intimately tied to disciplinary identities. Unsurprisingly, then, these “processes of disciplinary demarcation” are highly relevant in the establishment of new disciplines as they usually provide them with “a founding narrative and articulate core problems, general approaches and constitutive methods” (Powell et al. 2007: 5). Retrospective historiographic contextualization can reveal whether we can evaluate such processes as ultimately ‘successful’ according to different questions: Has the designation been (widely) adopted? Is the coining or adoption of the term perceived as having been influential in the formation of the discipline? Following Powell et al. (2007), reconstructions of such naming processes can be called ‘baptism stories’. This paper will trace multiple baptism stories for an odd yet interesting case of a discipline seemingly resistant to consistent naming: the study of writing.

Recent works published within the context of or addressing the study of writing often include or even commence with highlighting the coexistence of its many names. The following example is taken from Haralambous (2019: 151, emphasis in original):

There have been attempts to invent new terms: the author uses the term graphemics (‘graphématique’ in French) as a counterpart to phonology, others have proposed ‘graphonomy’, ‘grammatology’ (this term, originally introduced by Gelb (Gelb 1963) [...], became famous through Derrida’s homonymous book (Derrida 1967), which is more philosophical than linguistic), and at a higher level: ‘grapholinguistics’ (according to the German term *Schriftlinguistik*), etc.

---

1. This paper is dedicated to Christa Dürscheid. 20 years ago,\* her seminal textbook *Einführung in die Schriftlinguistik* (2002) was published. Often referred to simply as ‘die Schriftlinguistik’ in the Germanophone realm, it is a truly groundbreaking book that—in the course of its impressive five editions, the latest of which was published in 2016—not only helped constitute and ‘break the ground’ for a field devoted to the study of writing but has since also contributed tremendously in promoting it in the German-speaking linguistic community and beyond (an example being the book’s Korean translation published in 2007). Furthermore, it has considerably shaped me as well as my career trajectory as a (grapho)linguist, and it was a great honor to write a book on writing with Christa (*Writing systems and their use*, Meletis & Dürscheid 2022). Christa, congratulations and thank you! \*This paper was originally written and submitted in 2022.

The terms listed here are by no means nonce words; indeed, they have all been consciously introduced at some point in the literature published within the study of writing. None of them managed to prevail over the others, however, which is how they all remain—albeit with divergent frequencies of occurrence—in use until this day. They are tied to different contexts, sometimes also distinct (sub)disciplines, as well as academic cultures and traditions—and they all have their own baptism stories, even if these are, in the case of the study of writing, often unspectacular stories of introductions of terms without a lot of fuss. Looking at the manifold attempts at providing the study of writing with a name, scholars in the field apparently do not abide to what W. C. Watt (1994b: xii) urges—that “[n]omenclatural questions [...] should [...] detain us only in idle moments”. Proclaiming a name for a field that has yet to be firmly delimited and defined, even if some—including Watt—may interpret it as putting the cart before the horse, is not a decorative activity but a strategy obviously believed to contribute to a large degree to just that—establishment. Names matter indeed in that they are not hollow shells but “strategic tools” (de Chadarevian 2002: 206). As Powell et al. (2007: 26) generalize, “[d]isciplinary formation is so diverse and ongoing development so variable that names are one of the few factors capable of providing and maintaining disciplinary identity”. Speaking of disciplinary identity, what does it tell us, then, that no label for the study of writing has been unanimously accepted and widely adopted?

This paper is not primarily intended as a contribution to the broader analysis of the importance and effects of naming processes, which was fascinatingly outlined in a case study of four disciplines far removed from linguistics (namely genetics, molecular biology, genomics, and systems biology) by Powell et al. (2007). While the reconstruction of conditions surrounding the coining and adoption of different terms for the study of writing may also, down the road, be compared with baptism narratives in/of such unrelated disciplines, the main goal here is to shed light on an important part of a historiography of the study of writing, research on which remains sparse (cf. also Meletis in press). Crucially, knowledge of the history of a discipline including an “[u]nderstanding [of] how scientific activities use naming stories to achieve disciplinary stories is important not only for insight into the past” (Powell et al. 2007: 5) but can provide valuable insight going forward. As the contributions collected in the present proceedings of a grapholinguistic conference show, the study of writing is (on the verge of) thriving again. In this context, acknowledging that negotiating its name is not a recent activity and examining trends and challenges in previous baptism stories can, in the best case, be informative and instructive with respect to any future efforts in further establishing the field.

The paper is structured as follows: In Section 2, a selection of prominent names that have been proposed for the study of writing will be pre-

sented individually. This is followed by a synoptic discussion of central common threads in Section 3. A short programmatic outlook in Section 4 closes the paper.

## 2. The Candidates

In the following, prominent ‘candidate’ designations for the study of writing will be presented based on several questions including: Who invented or first used the term, and in which context? Was it then adopted by others, and why (not)? What is the term’s formal structure, i.e., which components does it consist of, what is their individual etymology and meaning, and what is their compositional meaning when combined? Conceptually, does the term suit the task of denoting the study of writing? Is it, for example, inclusive (enough), considering different perspectives on writing? What other, possibly non-writing-related meanings does the term have, and have these interfered with its use as a name for the study of writing? Note that the collection of terms included here is, of course, non-exhaustive. It is an ultimately subjective selection based on my own experience in and with the field and the literature that has been produced in it, and it is—even if this is attempted as best as possible—certainly not free from biases (concerning, for example, my own discipline or research community, cf. Meletis 2021a).

General trends and challenges characterizing attempts at naming the study of writing will already be mentioned throughout when a given term illustrates a common feature especially well; they will, however, be systematically collected in Section 3.

### 2.1. Grammarology: Gelb’s Ill-Fated Term

One of the first and most persistent designations for the study of writing is *grammarology*, a “modern formation from Gk γραμματο-, the combining form of γράμμα ‘letter’ and -λογία ‘teaching’” (Coulmas 1996a: 173). The first time it was more widely disseminated was in assyriologist Ignace J. Gelb’s *A study of writing* (1952),<sup>2</sup> a seminal book that ushered in a new era in the study of writing systems. Gelb’s adoption of the name was inspired not by previous uses—with different meanings—in German and French (cf., for example, Hasse 1792, Massé 1863) but by a different term, *grammatography*, found in the title of the English translation<sup>3</sup> of

2. Note that in this paper, the book’s second edition (published in 1963) is cited.

3. As Gelb (1963: 273, n. 46) himself notes, the German original of Ballhorn’s book does not use the term; it is titled *Alphabete orientalischer und occidentalischer Sprachen: zum Gebrauch für Schriftsetzer und Correctoren* (1847).



Friedrich Ballhorn's treatise of different 'alphabets of ancient and modern languages' (1861). Switching from *-graphy* to *-logy* makes sense, as concerning the field's scope, Gelb's aim was not a collective description of different writing systems merely for description's sake but to lay the foundation for an entire 'study of' writing.<sup>4</sup> In other words, Gelb's (1963: 23) intention was to contribute to the creation of a new field, and as is common in the course of this process, a potential name is provided: "The aim of this book is to lay a foundation for a full science of writing, yet to be written. To the new science we could give the name 'grammatology'." In the next sentence, he goes on to mention less suitable alternatives: "This term seems to me better suited than either 'graphology', which could lead to a misunderstanding, or 'philography' (a new term coined in contrast to 'philology'), which is not so exact as 'grammatology'" (Gelb 1963: 23). As will become apparent in the course of this paper, both of these operations are extremely common in the context of attempting to name the study of writing: scholars mentioning the novelty or unestablished status of the field and, in the same vein, arguing for their designation of choice while often listing the disadvantages of available alternatives.

The story of *grammatology* reveals yet another very common feature of the terminological history of the study of writing: drastically put, the 'derailing' of terms due to their use in other contexts and with divergent meanings. In the case of *grammatology*, this occurred very visibly and with lasting effects when French philosopher Jacques Derrida adopted—with acknowledgment (cf. also Daniels 1996a: 3)—the term for his influential and programmatic post-structuralist treatise *De la grammatologie* (1967, translated as *Of grammatology*, [1977] 1997).<sup>5</sup> While Derrida does focus on writing and its status, his *grammatology* is used in a "somewhat different though also related sense [...] to designate a theory of writing which he understands as a critique of the logocentrism of the Western intellectual tradition since Aristotle, which considers the sign (writing) as a mere supplement rather than an epistemic force in its own right" (Coulmas 1996a: 173). Interestingly, Derrida ([1977] 1967: 28, emphasis in original) also mentions other designations when describing his envisioned grammatology: "Graphematics or grammatography ought no longer to be presented as sciences; their goal should be exorbitant

---

4. Eckardt (1965: 4f.) criticizes also the other component of the term as restrictive: „Doch scheint mir auch diese Bezeichnung [= Grammatologie, DM] nicht ganz zufriedenstellend. Es handelt sich ja nicht um eine ‚Wissenschaft der Buchstaben‘—denn neben ‚Schrift‘ bedeutet γράμμα auch ‚Buchstabe‘—sondern um die Schrift in ihrer Gesamtheit.“ [“But even this designation [= grammatology, DM] seems to me not quite satisfactory. After all, it is not about a ‘science of letters’—for besides ‘writing’ γράμμα also means ‘letter’—but about writing in its entirety,” my translation].

5. Cf. Van de Mierop (2021) on Gelb's use of the term and Derrida's eventual appropriation.

when compared to *grammatological knowledge*.”<sup>6</sup> Not only does this echo the above-mentioned difference (in scope?) between *grammatography* and *grammatology*, but it also brings into play *graphematics* and reveals an awareness of this term.

Despite Derrida’s influential borrowing of the term, three decades later, in 1996, *grammatology* was still going strong, as is underlined by the publication of two books highly relevant to the study of writing. In his *Blackwell encyclopedia of writing systems*, linguist Florian Coulmas (1996a: xxv) writes: “No student of writing can dispense with the seminal works of Marcel Cohen, David Diringer, Ignace Gelb and Hans Jensen which have laid the groundwork for the scientific study of writing. More than 40 years ago Gelb proposed the term ‘grammatology’ for this field of inquiry.” In *The world’s writing systems* (cf. Daniels & Bright 1996), which to this day remains the most complete edited collection of descriptions covering a wide range of writing systems, one of the editors, Peter T. Daniels, who had already used *grammatology* in his earlier work (cf. Daniels 1990), observed that “[n]o name for this field of study has ever become widely accepted: ‘grammatology’, proposed in the mid twentieth century, is better than most” (1996b: 1). Crucially, both mentions of the term do not sweep under the rug its tentative nature as a ‘proposed’ term. Noteworthy is also Daniels’ (1996a: 3, emphasis in original) observation that *grammatology* “parallels *phonology* and *morphology*, the branches of linguistics that study sounds and meaningful units”; the reason this is interesting is that it tells us something about the intended scope of the field as well as its affiliation with—or even incorporation into—an established discipline (in this case linguistics), which are aspects closely tied to the proposal of names for fields of study. 1996 really was a remarkable year for the study of writing, as John Sören Pettersson also published his *Grammatological studies: Writing and its relation to speech*, an unfortunately little-received treatise addressing theoretical and methodological approaches to the subject of writing. More recently, *grammatology* is used only sporadically, e.g., by Zhong (2019)<sup>7</sup>, and the decline of occurrences in pertinent publications suggests that it may have been superseded by its alternatives—one of them being *graphonomy*.

---

6. The term *graphology* also features in his book (see Fleming 2016 and Section 2.3).

7. In her *Chinese grammatology: Script revolution and literary modernity, 1916–1958*, Yurou Zhong is not as much interested in a linguistic analysis of Chinese writing and Latinization efforts as in the fact that “the eventual retention of [Chinese] characters constituted an anti-ethnocentric, anti-imperial critique that coincided with post-war decolonization movements and predated the emergence of Deconstructionism” (<http://cup.columbia.edu/book/chinese-grammatology/9780231192637>, accessed November 2, 2022). This places her use of *grammatology* semantically somewhat between that of Gelb and Derrida, if a little closer to Derrida’s.

## 2.2. Graphonomy: Hockett's Little-Known Solution

In 2018, Peter T. Daniels' *An exploration of writing* was published, a monographic amalgamation of his decades-long research on writing systems that was—given his undeniable status as an authority in the field—long-awaited. The book's table of contents already foreshadows a terminological shift for Daniels, as its twelfth chapter is titled 'Graphonomy and linguistics'. This marks a change from *grammato-* to *grapho-*, deriving from Greek γράφω 'scratch, carve', as well as from *-logy* to *-nomy* from Greek νόμος 'law', which as a suffix signifies a system of rules, laws, or knowledge about a body of a particular field. Already in the book's introduction, Daniels (2018: 4f., emphasis in original) explains, in a footnote, why he now prefers *graphonomy* over *grammatology*:

The term [grammatology, DM] has become tainted in recent years: some scholars have taken it to refer to a school of writing-systems studies that holds to the Principle of Unidirectional Development<sup>8</sup> [...] and some other notions supported by Gelb; and the French philosopher Jacques Derrida borrowed it (with acknowledgment) to label a certain approach within Postmodern literary criticism. Therefore, I prefer 'graphonomy', which was introduced by Charles F. Hockett, [...] making explicit the analogy *astrology* : *astronomy* :: *graphology* : *graphonomy*.

He subsequently provides interesting details explaining why the "term could have been, but wasn't, popularized" (Daniels 2018: 5), including the fact that according to a handwritten note in one of Hockett's posthumously published manuscripts dealing with writing ('Speech and writing', 1952, published in 'Two lectures on writing', 2003), he had planned to define *graphonomy*—but ultimately did not. Ironically, a clear definition including a delimitation of the field's scope and aims is also missing from Daniels (2018) and subsequent works such as Daniels (2021), which even includes the term in its title ('Foundations of graphonomy').

What was likely detrimental to a larger dissemination of the term was the context of its introduction: Predating Gelb's use of *grammatology* by a hair, Hockett (1951) first mentions and discusses *graphonomy* in a review of John DeFrancis' book *Nationalism and language reform in China* (1950). The relevance of reviews notwithstanding, the attention they receive is arguably (and with exceptions) rather negligible when compared with that attracted by other types of publications, and in this particular case it is justified to rather drastically claim that Hockett's in-

---

8. This now-refuted principle propagated a teleological evolution of writing systems; Gelb (1963: 201) formulated it like this: "[...] in reaching its ultimate development writing [...] must pass through the stages of logography, syllabography, and alphabetography in this and no other order". Cf. for a discussion of counterevidence Daniels (2018: 133–135).

roduction of *graphonomy* was ‘buried’ in a review, and that this is likely the reason it never gained traction. Importantly, it is—as so often—not only the field’s designation that is discussed here, but also its breadth and relation to linguistics (and, in this case, also anthropology):

Books like De Francis’s—and reviews of them—will be easier to write when it is realized that the field of science primarily involved is not linguistics, but the yet unnamed study of writing and writing systems, and when at least some preliminary codification of the latter field has been done. Since the logical label for this sister-branch of anthropology, namely ‘graphology’, is otherwise occupied, let us follow the students of celestial phenomena in a removal to the suffix *-onomy*, and speak of GRAPHONOMY. Like other branches of anthropology, graphonomy has a pure and an applied angle; De Francis’ book involves both angles, but perhaps primarily the latter. Graphonomy can only progress on the basis of sound linguistics [...]. (Hockett 1951b: 445, emphasis in original)

While Hockett separates the “yet unnamed study of writing and writing systems” from linguistics, he later does relate the two by stating that graphonomy “can only progress on the basis of sound linguistics”. We will return to this complex relation—and question of the independence of the study of writing—in the discussion of grapholinguistics (Section 2.5) and general common threads (Section 3).

Another noteworthy use of the term came twenty years after Hockett’s review: computational linguist Sture Allén adopted the term in the title of his 1971 *Introduktion i grafonomi: Det lingvistiska skriftstudiet* (‘Introduction to graphonomy: The linguistic study of writing’). The fact that this was a Swedish-language publication makes this an appropriate point to emphasize another recurring aspect relevant in a discussion of attempts at naming the study of writing: introductions or uses of terms in languages other than English. As will be shown below for *Schriflinguistik*, the fact that terms may very well already be accepted and even widely established in other languages does not preclude a more international, English-speaking community from subjecting them to considerable scrutiny. Taking a closer look at the Swedish line of using *graphonomy*, Allén’s mentioned introduction was written in co-operation with Staffan Hellberg, who, in the subsequent publication of his English-language dissertation *Graphonomic rules in phonology: Studies in the expression component of Swedish* (1974), also relies on the term. The title alone (especially its inclusion of *phonology*) implies that Hellberg embeds graphonomy (as a phenomenon to be studied, as a field, or as both?) in a linguistic context. He fails at giving it a fixed meaning, however, as Wolfgang Börner notes in his review, which from a terminological perspective proves illuminating:

Hellberg verwendet weder den im Wortsinn normativen Terminus *orthography* noch den strukturalistisch vorbelasteten Namen *graphemics* (*graphology*



steht nicht zur Verfügung), sondern wie sein Lehrer Sture Allén den Terminus *graphonomy*. Dieser wandelt jedoch im Verlauf der theoretischen Diskussion seine Bedeutung. S. 1 wird *graphonomy* als autonome Schriftkomponente definiert: "The expression part of spoken language is often termed phonology. As its counterpart for written language, the term graphonomy has gained ground ...". Das Ziel der Arbeit ist die Untersuchung der "relation between phonology and graphonomy" (p. 1). Ein "graphonomic environment" (p. 45) ist folglich ein aus Buchstaben bestehender Kontext. Andererseits ist eine "graphonomic rule" (p. 42, 43 und passim) eine orthographische, d.h. Laut und Buchstaben verknüpfende Regel und in p. 201, Anm. 20 wird *graphonomy* auf einmal als "all (relevant) graphonomic rules," also als Äquivalent zur Orthographie vorgestellt. Noch mehr umfaßt *graphonomy* in p. 47: "exception features in the lexicon as well as the interspersed spelling rules". (Börner 1977: 337, emphasis in original)<sup>9</sup>

Not only does Börner (1977: 337) mention and contextualize other writing-related terms, distinguishing them from *graphonomy*, but in his critique it also becomes clear that Hellberg's use (or rather uses) of *graphonomy* is meant to designate primarily written structures (or certain features thereof, for which the adjectival form *graphonomic* is used), whereas Allén's book title had previously employed *graphonomy* at a meta-level, i.e., as the title of the study of writing. This, then, addresses a feature inherent in the majority of designations discussed in this paper: a subject-discipline ambiguity that is, however, not restricted to writing but widespread in linguistics (and many disciplines)—take *phonology* or *morphology*, levels of language and simultaneously disciplines studying them. Given the prominence of these latter terms, this polysemy usually does not stand in a way of a widespread dissemination, which means *graphonomy*'s non-success is likely rather based on the marginal status of writing as a research subject (especially in linguistics and especially at the times of Hockett and then also Allén) as well as the fact that works in which *graphonomy* was prominently used were little-received. It remains to be seen whether Daniels' recent (re)adoption of the term will lead to a reevaluation of its suitability and more widespread recognition.

---

9. "Hellberg uses neither the literally normative term *orthography* nor the structuralist-biased name *graphemics* (*graphology* is not available), but like his teacher Sture Allén the term *graphonomy*. However, this name changes its meaning in the course of the theoretical discussion. On p. 1 *graphonomy* is defined as an autonomous component of writing: 'The expression part of spoken language is often termed phonology. As its counterpart for written language, the term graphonomy has gained ground ...'. The aim of the paper is to investigate the 'relation between phonology and graphonomy' (p. 1). A 'graphonomic environment' (p. 45) is thus a context consisting of letters. On the other hand, a 'graphonomic rule' (p. 42, 43 and passim) is an orthographic rule, i.e., a rule linking sounds and letters, and in p. 201, note 20 *graphonomy* is suddenly presented as 'all (relevant) graphonomic rules', i.e., as equivalent to orthography. *Graphonomy* covers even more in p. 47: 'exception features in the lexicon as well as the interspersed spelling rules'" (my translation).

Before turning to the next candidate designation, other meanings of *graphonomy* shall be mentioned as they may also have contributed to a hesitance in using it. Firstly, it is close to a likewise writing-related term in which *-ics* replaces the *-y*: *graphonomics*, formally resembling *linguistics*, is “the multi-disciplinary field of fundamental and applied experimental research of handwriting and related skills” (taken from *graphonomics.net*, accessed October 19, 2022). The superficial and to some degree thematical closeness of *graphonomy* and *graphonomics* is undeniably not as severe as the complete collapse of two more drastically divergent meanings in the term *graphology* (see next section). Notably, in the view of semiotician W. C. Watt, who also published extensively on writing systems and edited the volume *Writing systems and cognition* (cf. Watt 1994a), the two related meanings of *graphonomy* and *graphonomics* apparently do collapse, as he notes: “There is no unified viewpoint from which to survey the study of writing systems. If there were, it could as well be called ‘graphonomics’ as anything else” (Watt 1994b: vii). In a later passage, he acknowledges the term’s above-mentioned non-linguistic origin, however, associating with it the advantage of not carrying any connotational baggage: “‘Graphonomics’ has gained currency through use by Kao, van Galen, and Hoosain (1986), and has the signal advantage of not being associated with quackery or dead grammatical theories. It parallels ‘linguistics’ in the broadest sense.” (Watt 1994b: xii, n. 1).

As for more strongly deviating meanings, while not as influential as Derrida’s appropriation of *grammatology* (but in spirit loosely related to it), *graphonomy*—specifically “Constitutive Graphonomy”—has in a different context been defined as “a post-colonial literary theory,” “the constitutive ethnography of writing systems” (Ashcroft 1989: 58). The fact that such uses in different contexts and with (more or less) new meanings and connotations occurred for both *grammatology* and *graphonomy* (and other terms as well, see below) highlights that there is no monopoly on using very general terms formed from semantically obvious and terminologically readily available elements such as *-graph-* and *-logy* or *-nomy*, which makes their repeated coining in varying disciplinary contexts understandable (and, from the perspective of each coining and coiner, justified). This is also the reason the use of the next candidate term as well as repeated attempts at reappropriating it are indeed quite relatable.

### 2.3. Graphology: Perfectly Parallel, but Already Occupied

The story of *graphology*, at least from the perspective of a forming study of writing in need of a name, is rather unfortunate. The obvious both formal and conceptual parallelism with *phonology* and *morphology* (see also Joyce 2023: 140), undeniably established and widely used linguistic

terms, can straightforwardly explain the motivation behind proposing *graphology* as the name for their written equivalent. According to German linguist Konrad Ehlich (2007: 728), this leaning on successful pre-existing terms is a symptom of a general terminological trend in the linguistic treatment of writing: “Die Terminologisierung [in der linguistischen Schriftforschung, DM] ist Ausdruck eines Teilhabeversuches am Nutzen dessen, was in der Phonologie mit einem ziemlichen Erfolg erreicht worden war.”<sup>10</sup> However, when the point in linguistics had been reached in which the subbranch dealing with writing had matured enough to require (or justify) a name of its own, *graphology* had already been taken—or, somewhat more drastically put, ‘derailed’—by “[t]he study of handwriting from the point of view of diagnostic psychology,” the basic assumption of which “is that features of handwriting [...] are indicative of character and personality traits” (Coulmas 1996a: 178). The disputed (pseudo-)scientific status of such a psychological handwriting-focused graphology (vs. uncontroversially accepted forensic handwriting analysis, which must be carefully separated from it),<sup>11</sup> which became popular at the end of the 19th century with works such as Klages’ (1917) *Handschrift und Charakter* (‘Handwriting and character’), shall not be discussed here. It is noteworthy, however, that it is often heavily scrutinized in linguistic works on writing (such as in Dürscheid 2016: 201f., n. 166).

Of relevance in the present historiographic account of terminology is that despite its dominant different meaning, “[s]ometimes the term ‘graphology’ is also used in analogy with ‘phonology’, that is, in the sense of *graphemics*” (Coulmas 1996a: 178; for *graphem(at)ics*, see next section). In this context, at least three main strategies of dealing with the term *graphology* need to be distinguished: (i) it is used in a linguistic reading without reference to its existing psychological meaning—either as a name for a linguistic phenomenon (i.e., a written module of language) or as a name of the field studying it, reproducing the above-mentioned ambiguity, (ii) it is rejected on grounds of its psychological meaning, or (iii) this meaning is acknowledged, but the term is reappropriated in the context of linguistics.

---

10. “Terminologization [in linguistic writing research, DM] is an expression of an attempt to share in the benefits of what had been achieved with a fair amount of success in phonology“ (my translation). Cf. also Wales (2014: 194, emphasis in original): “From Gk *graphos* ‘written’, linguistics has spawned a whole set of terms to do with the study of written language, most by analogy with the study of speech in PHONETICS and PHONOLOGY.”

11. This perceived pseudo-scientific status is something *graphology* shares with the terminologically parallel *astrology*.

When searching for adoptions of the term in linguistic publications,<sup>12</sup> quite a few can be found—both in noun form (*graphology*) and in adjectival form (*graphological*).<sup>13</sup> Examples include Logan (1973: Chapter III), who, in his study, devotes an entire chapter to ‘graphology’ (parallel to another chapter on ‘phonology’); he defines it as a synonym of ‘writing system’ (Logan 1973: 32) and mentions that he adopted the term from McIntosh’s (1961) ‘Graphology and Meaning’ (Logan 1973: 32, n. 1). Indeed, linguist Angus McIntosh is claimed to have been one of the first to use *graphology* systematically in this linguistic reading, as also outlined—and later contextualized with respect to the non-linguistic meaning of the term—by Gómez-Jiménez (2015: 71, emphasis in original):

*Graphology* is a linguistic level of analysis that comprises the study of graphic aspects of language. This term was first brought into use in linguistic studies in the sixties by McIntosh (1961), who considered it an analogous mode to that of phonology. In his paper ‘Graphology and Meaning’, he declared he had used graphology ‘in a sense which is intended to answer, in the realm of written language, to that of ‘phonology’ in the realm of spoken language’ (1961: 107).

Slightly later, well-known British linguist David Crystal started using the term, first together with Derek Davy (cf. Crystal & Davy [1969] 1979) and then in many later publications (such as Crystal 1980: 168f., [1987] 1997: 184–209, 2003: 210f.; cf. also Spitzmüller 2013: 111f. for a discussion of Crystal’s use of the term). One of his definitions reads: “Graphology, coined on analogy with *phonology*, is the study of the linguistic contrasts that writing systems convey” (Crystal [1987] 1997: 187, emphasis in original). As both McIntosh’s and Crystal’s uses of the term show, the pre-existing and more prominent psychological meaning is not always mentioned for clarification, even if it can be assumed that the authors were, of course, aware of it. In the majority of works, however, such a delimitation is practiced, an example being Wales’ (2014: 194, emphasis in original) *Dictionary of stylistics*, where *graphology* is defined as follows:

“The study of such units in a language [graphemes and allographs, DM] is called *graphemics*, or *graphology*. (In popular usage *graphology* also

12. Notably, what I carried out here were simple searches on Google Books and Google Scholar and not sophisticated and in-depth literature searches, which would likely yield more interesting results.

13. One slightly deviating form can be found in Louis Hjelmslev’s (1947: 69, my emphasis) ‘Structural analysis of language’, where he uses *graphbiology*—although it is not clear whether this may be a typo: “Thus, Saussure would have it that the sounds of a spoken language, or the characters of a written language, should be described, not primarily in terms of phonetics or of *graphbiology*, respectively, but in terms of mutual relations only, and, similarly, the units of the linguistic content (the units of meaning) should be described primarily not in terms of semantics but in terms of mutual relations only.”

refers confusingly to the study of handwriting as a means of character analysis).” She goes on to mention that “[g]raphology can also refer to the writing system of a language, as manifested in handwriting and typography; and to the other related features [...], e.g., capitalization and punctuation.”

In most works in which the name of the study of writing is addressed explicitly, the unsuitability of *graphology* is pointed out (cf., for example, Hockett 1951b: 445; Gelb 1963: 23; Nerius 1986: 38; Haralambous 2019: 151), occasionally with an explicit mention that it “is otherwise occupied” (Hockett 1951b: 445) and “could lead to a misunderstanding” (Gelb 1963: 23), such as by Daniels (2018: 5), who states (in parentheses, and rather critically) that “[g]raphology is the pseudoscience of diving someone’s personality from their handwriting”. Interestingly, in some of these passages, often between the lines, not only a slight annoyance with the term’s prior occupation but also a related (implicit) lamenting can be perceived. Watt (1994b: xii), for example, who approaches the study of writing from a cognitive rather than a purely linguistic perspective, writes that “[t]he ideal analog of ‘phonology’ would be ‘graphology’, the study of individual letter-components of a writing-system (both studies would then deal with elements nicely fissionable into distinctive features [...]); but it remains to be seen whether this term can be freed of its previous associations”. It is words and phrases such as ‘ideal’ and ‘can be freed’ that convey a sense of regret that *graphology* is unavailable.

Konrad Ehlich, a scholar of writing instrumental in shaping the German grapholinguistic tradition (see Section 2.5), wanted to reappropriate the term after acknowledging that its predominant meaning is a different one (cf. Ehlich 2001: 63):

The term ‘phonology’ uses the affix ‘-logy’, and in doing so, it makes reference to the inner systematic quality of the phoneme system. I think, it is worthwhile to keep this line of thinking in the case of graphics. So I would like to propose re-introducing the term ‘graphology’ into the theoretical framework, as a systematically founded term. Graphology in this sense is no longer a term referring only to expression characteristics of individuals, but it is a term which refers to the inherent organized structure of writing. (Ehlich 2001: 65)

What is noteworthy about Ehlich’s attempt at reintroducing *graphology* is the specific meaning tied to it. It does not correspond completely with different prior uses that can be considered mostly synonymous with *graphem(at)ics* or ‘writing system’ (see below) but is intended to underline the internal functional organization of writing, which, crucially, includes its oft-neglected materiality. In other words, the term “highlights that the material subsystem of writing has its own systematicity. What Ehlich means by ‘systematicity’ is the fact that writing is spatially organized in a way that allows studying it as a visual system

completely without the consideration of linguistic facts” (Meletis 2020: 34). Ehlich’s reading of the term, despite its fine-grained sophistication, was never widely adopted.

Finally, and somewhat humorously, *graphology* was also appropriated by a more philosophical tradition, by Juliet Fleming (2016) in her book *Cultural graphology: Writing after Derrida*. The reason this is humorous is that, as the title suggests, this use of *graphology* follows in the direct footsteps of Derrida’s adoption of *grammatology* and also somewhat resembles the above-mentioned appropriation of *graphonomy* in the context of cultural studies. In the book’s introduction, titled ‘From Grammatology to Cultural Graphology’, Fleming (2016: 1) writes: “Cultural graphology names a new approach to the study of texts” and contextualizes it—following Derrida’s own (vague) ideas about a cultural graphology—within the field of book history.<sup>14</sup> A straightforward definition of cultural graphology is not (and possibly cannot be) given but must be deduced from passages such as this:

Another name for this discipline, which would combine (at the very least) psychoanalysis, literary history, bibliography, book history, the sociology of texts, and information technology, is, of course, cultural graphology. (Fleming 2016: 39)

#### 2.4. Graphem(at)ics, Orthography, Writing Systems Research: Fitting but Restricted

The next candidate in some ways parallels *grammatology*, *graphonomy*, and *graphology*, and in other ways it does not. *Graphemics*, or its longer form *graphematics*, which are found in many languages (German *Graphemik/Graphematik*, French *graphémique/graphématique*, Spanish *grafémica/grafemática*, Italian *grafemica/grafematica*, Swedish *grafemik/grafematik*, etc.), again denote both a part of a language system—its functional written component (sometimes distinguished from *graphetics*, its material component)—and, as with the other above-mentioned terms, the field devoted to analyzing said component.

What needs to be clarified first with respect to this term is whether there exists a semantic difference between its shorter version *graphemics* and the longer *graphematics*, both of which are modelled after speech-related linguistic fields (*phonemics* and *phonematics*, which are most often

---

14. More specifically, she attempts a deconstruction of said field: “[...] we can use the resources of deconstruction to shake up and enlarge the field that, for the time being, and in spite of its obvious limitations, might still be called book history” (Fleming 2016: 16, emphasis in original).

considered synonymous). Usually, they are treated as equivalents, making the choice between them a matter of taste; however, a slight preference for *graphemics* can be observed in research with an Angloamerican origin, while *graphematics* (both as an English term and its translations into other languages) is more common in research stemming from other scholarly traditions such as the German one.<sup>15</sup> It is only in exceptions that a fine-grained difference is intended by the two terms: In their German textbook, for example, Fuhrhop & Peters (2013: 203, emphasis in original) use the associated adjectives to highlight a conceptual distinction:

'*Graphemisch*' wird hier verwendet, weil der direkte Bezug zum 'Graphem' hergestellt wird; '*graphematisch*' hingegen bezieht sich auf die gesamte Graphematik, als grammatisches Teilsystem.<sup>16</sup>

As for the term's history, according to Piirainen (1986: 97), the "theory of graphemics was founded in 1930's [sic] by the linguistic schools of Prague and Helsinki"; cf. also Coulmas (1996a: 176): "The case for an autonomous graphemics has been made most forcefully and consistently since the 1930s by members of the linguistic school of Prague." While the Prague school—and most vocally its member Josef Vachek—was instrumental in the theoretical establishment of a linguistic graphemics, the focus here shall remain on the terminological side of this process. Here, what is interesting in the case of *graphem(at)ics* is that its first coining or use likely happened without much ado due to the exact—and therefore obvious—terminological "parallelism of phonemics and graphemics" (Pulgram 1951: 19); cf. also Hall (1960: 13, emphasis in original): "In recent years, following upon the development of phonemic theory, there have been several discussions of the relation of phonemes to their written notation, and parallel to *phoneme* and *phonemics*, the terms *grapheme* and *graphemics* have come into use." *Graphemics*, in other words, was simply a natural choice for the linguistic subfield (and sublevel) concerned with units of writing, so whoever used it first likely did not sell its adoption as an inventive achievement. Also, unlike *graphology*, it was not already taken by an altogether different field (see above). It is likely for these reasons that early uses of *graphemics* do without elaborate (or

15. An interesting illustration of this English vs. non-English correlation of the shorter and longer versions is the name of the 2018 iteration of the /*gʁafematik*/ conference series, which was called *Graphemics in the 21st Century* (cf. <http://conferences.telecom-bretagne.eu/grafematik/>, accessed November 1, 2022). Here, French *graphématique* is the equivalent to English *graphemics* (cf. also Haralambous 2019: 151) although there exist respective correspondences in both languages (English *graphematics*, French *graphémique*).

16. "'Graphemic' is used here because direct reference is made to 'grapheme'; 'graphematic', on the other hand, refers to the whole graphematics as a grammatical subsystem" (my translation).

sometimes any) definitions,<sup>17</sup> and the scope and tasks of the designated field are only at times characterized (cf., for example, Bazell 1956).

It was the German(ist) research tradition and community that adopted *graphemics* for the study of the specifically linguistic functions of writing (in a narrow sense), encompassing aspects such as a grapheme definition, allography, and graphotactics, and it has since consistently stuck with the term—albeit, as mentioned above, mostly in its longer form *graphematics*.<sup>18</sup> Crucially, even when *Schriftlinguistik* as a designation for a broader, more interdisciplinary study of writing (see below) had not yet been established, *graphematics* was not intended to fill that void but was predominantly used with its specific meaning alongside other terms such as *graphetics* and *orthography* (cf., for example, Augst 1985; Gallmann 1985; Günther 1988; Fuhrhop & Peters 2013; Berg & Evertz 2018; Berg 2019). In other words, in the German reading, *graphematics* does not denote the multifaceted study of writing in its entirety but indeed only the linguistic part of it—and possibly not even all of that, either. In Dürscheid's seminal *Einführung in die Schriftlinguistik* (2002), for example, graphematics was treated in an eponymous chapter alongside chapters covering, among others, the history of writing, literacy acquisition, and orthography. Especially the coexistence of dedicated chapters on graphematics and orthography must be commented on, both because it insinuates that they are not the same phenomenon and because the latter, like graphematics, has also been (and is partially still being) used as a *pars pro toto* designation for the linguistic study of writing, especially in the Angloamerican realm.

This is not the place to discuss in detail how in English-language works published mostly by scholars socialized in an English-writing culture, *orthography* (from Greek *ὀρθο-* 'correct', coupled with the recurring *-graphy*) is used in a descriptive reading related in sense to the above-mentioned *graphematics* or even the broader *writing system* (see below). In short, the reason for this could be that for varieties of written English, no binding orthographic codification regulated by an official authority of linguistic policy exists—as it does for the German writing system with the *Amtliche Regelung* issued by the *Council for German Orthography* (cf., for more details on the difference between descriptive and prescriptive meanings of *orthography*, Meletis 2021a; Meletis & Dürscheid 2022: Chapter 5). While in Germanist research, *graphematics* and *orthography* thus de-

---

17. Cf. Hamp (1959: 1), who, in a paper titled 'Graphemics and paragraphemics' (!), writes: "It is not the purpose of the present note to discuss graphemics in any detail; nor is graphemics as such the central theme."

18. Notably, Althaus' (1980) article in a German-language linguistic lexicon was still titled 'Graphemik', so the shorter version was also used in German before becoming dispreferred.



note different phenomena,<sup>19</sup> in literature with an Angloamerican origin, *orthography* is frequently used in a more general manner so that, for example, Richard Venezky's (1970) seminal book on the English writing system (and not just its normative aspects) is called *The structure of English orthography*. And for *orthography*, too, we encounter the typical ambiguity, as to this day, it is used also for the enterprise of studying orthographic (or graphematic) structures, as in Condorelli's (2022) *Introduction to historical orthography* (cf. also Condorelli 2020), in which it is defined as "the scientific study of writing in history" which "focuses on the description and study of orthographies, their development over time, as well as the forces and the processes which shaped and directed modifications in historical writing features" (Condorelli 2022: 3).<sup>20</sup>

In some modern works fundamentally based on earlier structuralist German research on writing, graphematics and orthography are seen as individual—albeit interacting and overlapping—components or 'modules' of a writing system, which itself is defined as the graphic and linguistic notation of a specific language. This view is most pronounced in Martin Neef's (2005, 2015) multimodular theory of writing systems originally devised for German and later other alphabets (cf. Meletis 2020 for a broader adaptation considering also non-alphabetic systems). It is this use of the term and concept of *writing system* that serves as a fitting transition to the final candidate designation that shall be mentioned in this section, the umbrella term *writing systems research*. It is, first and foremost, the title of a Taylor & Francis journal that was published from 2009 to 2019, when it was, unfortunately, ceased. Rarely, the term can also be found in individual publications such as Mark Sebba's (2009) 'So-

---

19. This is also evident in the title of Gerhard Augst's (1986) edited volume *New trends in graphemics and orthography*.

20. It must be noted both that Condorelli (2022: Chapter 1) is aware of and does discuss the different meanings of *orthography* and that there are, of course, orthography-like normative phenomena also in historical stages of writing systems that are not officially regulated (cf., for example, Mihm 2016). Also, as concerns the historical study of writing systems, a different tradition rooted mostly in German-language research must be mentioned, which goes by *historical graphematics* (cf., for instance, Elmentaler 2018). On the webpage of the book series *LautSchriftSprache* (Reichert Verlag), which is associated with the eponymous conference series focusing on the diachronic study of writing, historical graphematics is defined as follows: "Als ein multidisziplinäres Forschungsgebiet stellt die [historische, DM] Graphematik die Brücke zwischen Philologie, Sprachgeschichte, Epigraphik und Semiotik dar. Daher beschreibt die historische Graphematik die allgemeinen Strukturen überlieferter Schreibsysteme" (cf. [https://reichert-verlag.de/buchreihen/sprachwissenschaft\\_reihen/sprachwissenschaft\\_lauteschriftsprache\\_scriptandsound](https://reichert-verlag.de/buchreihen/sprachwissenschaft_reihen/sprachwissenschaft_lauteschriftsprache_scriptandsound), accessed November 1, 2022). ["As a multidisciplinary field of research, [historical, DM] graphematics represents the bridge between philology, language history, epigraphy, and semiotics. Therefore, historical graphematics describes the general structures of recorded writing systems."]

ciolinguistic approaches to writing systems research', in Joyce & Meletis (2021), where it is given preference over *grapholinguistics*, which there is mentioned as its synonym, or in Joyce (2023). *Writing systems research* has the obvious benefits of being rather neutral and broad; when looking at the aims and scope of the now-defunct journal, for example, a multidisciplinary yet curiously selective picture is drawn of what the associated field could cover; it is reproduced in the following.<sup>21</sup>

---

*Writing Systems Research* (WSR) publishes work concerned with any issue to do with the analysis, use and acquisition of writing systems (WSs) such as:

1. The linguistic analysis of writing systems at various levels (e.g., orthography, punctuation, typography), including comparative WS research.
  2. The learning and use of writing systems, including:
    - Learning to read and write in children (normal and disabled children, bilingual children acquiring two WSs, deaf children) and adults (illiterates, learners of second language WSs).
    - The psycholinguistic processes of reading (grapheme recognition, word recognition) and writing (spelling, handwriting) in specific writing systems and in cross-orthographic comparisons.
  3. Neurolinguistics and writing systems (e.g., lateralisation, reading pathologies, reading and writing disorders).
  4. The correlates of writing systems:
    - Writing systems and metalinguistic awareness (e.g., phonemic awareness, word awareness).
    - Cognitive consequences of writing systems (e.g., visual memory, representations of time sequences).
  5. Writing systems and computer/new media:
    - Computers in reading and writing.
    - Consequences of computers/new media on writing systems and their use.
    - Computer modelling of writing systems.
- 

This list reveals the journal's (and field's?) linguistic and psychological/psycholinguistic as well as cognitive focus and mentions—somewhat out of place—also 'computers' and 'new media' (rather than the broader 'technology') as an additional perspective on writing systems. What is strikingly omitted is the sociolinguistic perspectives that had been characterized by Sebba (2009) in his article published in the journal's inaugural volume. The journal thus sees literacy practices and in general the use of writing systems mainly from a processing perspective, not a more user-oriented communicative one.

Furthermore, the specific use of 'writing system' rather than just 'writing' (or *Written Language and Literacy*, which is the title of another

---

21. <https://www.tandfonline.com/action/journalInformation?show=aimsScope&journalCode=pwsr20> (accessed November 1, 2022).

writing-related journal, published by John Benjamins) implies a field that is more restricted than a comprehensive study of writing, as is also argued in Meletis (2020: 3, n. 3, emphasis in original):

although its focus on writing systems is obviously justified, the term insinuates a narrower scope than what is actually studied by grapholinguistics: for example, solely graphetic research endeavors, such as studies that test which connotations or emotions different typefaces evoke, are definitely grapholinguistic but not about the writing system *per se*. Such questions might not always be seen as writing systems research.

In this quote, the designation of choice for the study of writing is *grapholinguistics*, to which we turn next.

## 2.5. *Schriftlinguistik*/grapholinguistics: A Question of Disciplinary (In)dependence and Tradition

We thus arrive at the nowadays most widely adopted—but by no means unanimously accepted—designation for the study of writing, *grapholinguistics*, and its relation to its widespread German sister term *Schriftlinguistik*. Although, in the meaning relevant here, *grapholinguistics* entered the Anglophone research realm only recently (through, among others, Neef's above-mentioned 2015 article 'Writing systems as modular objects: Proposals for theory design in grapholinguistics'), its history is a much longer one. In German, *Schriftlinguistik* (and its synonym *Grapholinguistik*) had been used since roughly 1980, at first mainly by the *Forschungsgruppe Orthographie*, a research group surrounding German linguist Dieter Nerius (cf. Nerius 2012), who is sometimes mentioned as the founder of the term (cf. Neef 2021; Dürscheid 2016: 12, n. 2). One of its first uses in print can be traced to 1986,<sup>22</sup> when Nerius used it in an article addressing concepts in the field of written language ('Zur Begriffsbestimmung im Bereich der geschriebenen Sprache'):

Diese Ansätze einer Linguistik der [geschriebenen Sprache] und einer Linguistik der [gesprochenen Sprache] oder, wie wir auch sagen können, einer Grapholinguistik und einer Phonolinguistik, gilt es weiterzuentwickeln und auszubauen. Für die Grapholinguistik, die hier im Mittelpunkt unseres Interesses steht, gehört dazu nicht nur die Untersuchung des Graphemsystems und der anderen graphischen Formeinheiten, [...] sondern auch die Untersuchung graphomorphologischer, grapholexikalischer, graphosyntaktischer, graphotextualer und natürlich auch graphostilistischer Erscheinungen, im weiteren Sinne also sowohl das System der [geschriebenen Sprache]

---

22. Neef (2021) notes that German linguist Helmut Glück had already used *Schriftlinguistik* in his habilitation thesis which was accepted in 1984 and published in 1987 (cf. Glück 1987: 13, 59).

als auch ihre Verwendung in der schriftlichen Kommunikation. (Nerius 1986: 37)<sup>23</sup>

Nerius does not provide a detailed definition but characterizes *Grapholinguistik* as ‘the linguistics of written language’ encompassing the study of both the system of written language—at various linguistic levels such as the ‘graphomorphological’ one—and its use in written communication. Given that the German-language journal in which his article was published also includes abstracts in English, Russian, and French for all its articles, translations of the term are provided: English *grapholinguistics*, Russian *графолингвистика* (‘grafolinguvistika’), French *grapholinguistique*.<sup>24</sup> From this, one can conclude that English *grapholinguistics* was Nerius’ translation of choice—or at least one he likely approved of. Indeed, *grapholinguistics* is a straightforward and uncontroversial translation of German *Grapholinguistik*. Interestingly, however, the latter is not the German term that would eventually prevail and become established. Shortly after Nerius’ article, in 1988, a volume co-edited by him and fellow German linguist Gerhard Augst already had the alternative *Schriftlinguistik* in its subtitle (cf. Nerius & Augst 1988). In the volume’s introduction, in commenting on writing-related works that had been published up until that point, Nerius (1988: 1) remarks: “Solche Arbeiten dokumentieren das Interesse der internationalen Linguistik an diesem Forschungsgegenstand und zeigen, daß sich hier eine eigenständige linguistische Teildisziplin, die Schriftlinguistik oder Grapholinguistik, entwickelt hat.”<sup>25</sup> This quote is relevant for two reasons that shall be addressed in more detail in the following: firstly, and terminologically,

---

23. “These approaches of a linguistics of [written language] and a linguistics of [spoken language] or, as we can also say, a grapholinguistics and a phonolinguistics, need to be further developed and expanded. For grapholinguistics, which is the focus of our interest here, this includes not only the study of the grapheme system and the other graphic form units, [...] but also the study of graphomorphological, grapholexical, graphosyntactic, graphotextual, and, of course, graphostylistic phenomena, in the broader sense, that is, both the system of [written language] and its use in written communication” (my translation).

24. In this context, the Croatian grapholinguistic tradition shall also be mentioned, whose most prominent representative is Mateo Žagar. In his research, which includes the 2007 book *Grafolinguvistika srednjovjekovnih tekstova* (‘Grapholinguistics of medieval texts’), he—with reference to Christa Dürscheid’s work (see below)—applies a grapholinguistic framework to historical texts. Cf. also Žagar (2020: 180): “With the introduction of modern, primarily structuralist, grapholinguistics, scholars can now work on a solid framework within which phenomena representing the distinct written realization of a linguistic unit are placed, together with the visual surroundings that optimise the transmission of a textual linguistic message [...]”.

25. “Such works document the interest of international linguistics in this research subject and show that an independent linguistic subdiscipline, [Schriftlinguistik] or [Grapholinguistik], has developed here” (my translation).

it marked the first step in *Grapholinguistik* being relegated to the status of a (mere) synonym of the preferred *Schriftlinguistik*; secondly, and more importantly, at the conceptual level, Nerius defines the field as a branch or subdiscipline of linguistics—albeit an explicitly ‘independent’ one.

As for the first of these points, the mentioned volume was just the initial step in promoting *Schriftlinguistik* as the new designation for the field. In 1993, the first edition of a now well-known German linguistic dictionary, the *Metzler Lexikon Sprache*, edited by Helmut Glück (cf. Glück 1993), included an entry ‘Schriftlinguistik’, and in 1995, a Festschrift for Dieter Nerius was published (cf. Ewald & Sommerfeldt 1995) which highlighted the term very prominently in its title *Beiträge zur Schriftlinguistik* (‘Contributions to *Schriftlinguistik*’). The arguably decisive moment in the term’s establishment, however, came with the publication of the first edition of Christa Dürscheid’s *Einführung in die Schriftlinguistik* in 2002. While, given the examples above, it was not the first book to carry the term in its title, it was not a collection of different shorter contributions to the field but a coherent single-authored textbook giving an overview of the field’s different facets, thereby systematically characterizing and arguably in large part constituting it in the first place. Interestingly, although by the early 2000’s, as outlined above, the term had already circulated for some time in the Germanophone linguistic community, Dürscheid wrote:

In diesem Buch wird der Standpunkt vertreten, dass die Schrift genuin ein Gegenstand der Sprachwissenschaft ist. Um dies kenntlich zu machen, trägt das Buch den Titel ‘Einführung in die Schriftlinguistik’, obwohl der Terminus ‘Schriftlinguistik’ bis heute nicht in den fachsprachlichen Gebrauch eingegangen ist. (Dürscheid 2016: 11)<sup>26</sup>

The perception at the time the textbook was written was evidently that although *Schriftlinguistik* was being used in specialized circles, it—as well as the field it is meant to label—had not yet been accepted into the canon of linguistics at large (see also below). This, notably, is something that Dürscheid’s textbook has managed to change following its publication. In 2012/13, German linguists Martin Neef and Rüdiger Weingarten (later also joined by Said Sahel) began editing a dictionary called *Schriftlinguistik* in the De Gruyter series *Dictionaries of Linguistics and Communication Science*, a companion series to the influential handbook series *Handbooks of Linguistics and Communication Science*. In the latter, the two-volume interdisciplinary handbook *Schrift and Schriftlichkeit/Writing and its use* edited by Hartmut Günther and Otto Ludwig (1994/1996) had been

---

26. “This book argues that writing is a genuine subject of linguistics. To make this clear, the book is entitled ‘Introduction to [Schriftlinguistik]’, although the term ‘Schriftlinguistik’ has not yet entered linguistic jargon” (my translation). Note that this passage is still intact in the textbook’s fifth edition published in 2016.

published, which, strikingly, did not utilize the term *Schriftlinguistik* that would eventually be picked as the title of the sister dictionary. For what follows, it is crucial to note that the dictionary series was meant (at least initially) to be bilingual; while the German versions of the dictionaries, the first of which was published in print in 2021,<sup>27</sup> include English definitions for all lemmas, the plan was to also publish entire equivalent dictionaries in English. Importantly, now, for the *Schriftlinguistik* dictionary, *Grapholinguistics* was chosen as the title.<sup>28</sup> This represented a vital step in establishing *grapholinguistics* not merely as an apparent translation of the superseded and little-used German *Grapholinguistik*—which in the dictionary itself is also treated as a mere synonym of *Schriftlinguistik* (cf. Neef 2021)—but to establish it officially and visibly as the English designation of a field that, in the German-speaking area, had already found a considerable footing. Note, however, that the *grapbo-* in *grapholinguistics* was by no means an obvious choice from a purely formal perspective, and certainly not an inevitable one.<sup>29</sup>

Take *Korpuslinguistik*, for example, which in English is *corpus linguistics*, or *Kontaktlinguistik*, which in English is *contact linguistics* (or sometimes simply referred to by the phenomenon studied, *language contact*). These German labels, now, are categorically different from words like *Psycholinguistik*, *Soziolinguistik*, and also *Grapholinguistik*, in which bound lexemes are combined with *-linguistik* (in English, too, they are bound: *psycholinguistics*, *sociolinguistics*, *grapholinguistics*), as *Korpus*, *Kontakt*, and also *Schrift* are all free lexemes. Accordingly, a two-part English translation of *Schriftlinguistik* following the pattern of *corpus linguistics* would have been a possibility, raising the question of which word would be the best English choice for the broad *Schrift*: *writing*, which itself is polysemous as it designates—among many other things—both the act of writing and the resulting product, producing the awkward-sounding *writing linguistics*? Or maybe the Latin-derived *script* (which thus more elegantly aligns with likewise Latin-derived *linguistics*)?<sup>30</sup> Indeed, *script linguistics* has been used sporadically (cf., for example, Rössler, Besl & Saller 2021: XXVI);

27. See <https://www.degruyter.com/serial/wsk-b/html/#volumes> (accessed November 2, 2022).

28. See <https://www.wsk.fau.de/baende/englischsprachige-wsk-baende/> (accessed October 24, 2022).

29. Notably, arguing about the *grapbo-* as the first (and obvious) constituent that recurs throughout the terminology used in the study of writing may be beside the point here and thus merely a cosmetic terminological analysis as the term's component that people actually appear to have a problem with is evidently *-linguistics*, which is interpreted as limiting the field's scope to linguistic questions (see below).

30. The mixture of Greek *grapbo-* with Latin *-linguistics* has indeed been criticized (Peter Daniels, pers. comm., Nov. 2020); see also a comment by user 'Coby Lubliner' under the blog entry 'Grapholinguistics' in the *Language Log* (<https://languagelog.ldc.upenn.edu/nll/?p=46324>, accessed November 2, 2022). Interestingly, for other

its core drawbacks are that *script* itself has been used with myriad different definitions, and these generally also have a narrower semantic scope than *writing* (see also the discussion of *scriptology* in the next section).

That *Schriftlinguistik* belongs to the free morpheme group while *grapholinguistics* is part of the bound morpheme group is not trivial but associated with an important semantic difference: the free morphemes in these designations stand for what is being studied by the respective fields: language contact, corpora, writing. By contrast, the bound morphemes are abbreviations for fields themselves (and associated methods, theories, paradigms, etc.). One of the criticisms that have been voiced against *grapholinguistics* is that as a designation, it evokes the latter group while the field that is in need of a name—the ‘study of writing’—is actually of the former type. Unlike *psycholinguistics* or *sociolinguistics*, thus, *grapholinguistics* is not the merging of two disciplines: when *psycho-* stands for *psychology* and *socio-* for *sociology*, what does the *grapho-* stand for? The sobering answer: A discipline that does not exist, a discipline that is—as this paper shows—assigned many names, for which *grapholinguistics*, in its entirety, as an attempt to translate the uncontroversial *Schriftlinguistik*, is admittedly a less-than-ideal workaround that is *not*—as claimed in Meletis (2020: 8)—exactly parallel to labels for other subfields of applied linguistics. Daniel Harbour (pers. comm., Oct. 2022) explains with regard to *grapholinguistics*:

It cuts the world up in the wrong way. We already have formal linguistics, neurolinguistics, psycholinguistics, sociolinguistics, etc. There is of course a degree of overlap between these (a sociolinguist can take a historical perspective and so end up doing sociohistorical linguistics; or sociophonetics; and a theoretical explanation can be given to some sociolinguistic variation). But for the most part, these subfields are distinct as to methods and subject matter. ‘Grapholinguistics’, qua term, gets the wrong end of the stick. Grapholinguistics does not sit alongside these areas as a separate subdiscipline. It cross-cuts them. Neurolinguistics and psycholinguistics rely heavily on, and feed significantly into, the study of writing systems. Written language is just as suited to sociolinguistic study as spoken language is. Historical linguistic methods likewise.

This very clearly reiterates that the decision of how to name the field is not merely a terminological one but one that feeds into the crucial questions of how the field is conceived and contextualized, what it covers, and what its boundaries are. As outlined above, Nerius (1986) had considered *grapholinguistics* a linguistic subdiscipline but had added that it was ‘independent’. What does this mean? It is likely related to Harbour’s reservations about *grapholinguistics*: the study of writing is inherently interdisciplinary and characterized by the adoption of multiple

---

etymologically (mostly) parallel designations—such as *psycholinguistics*—this mixture does not appear to be a problem.

perspectives. Placing *grapholinguistics* alongside *psycholinguistics* and *sociolinguistics*, now, means it is separated from them although grapholinguistics has psycholinguistic and sociolinguistic questions at its core, and simultaneously, psycholinguistics and sociolinguistics deal with writing, too.<sup>31</sup>

The second major criticism voiced against *grapholinguistics* is that the interdisciplinarity needed to study the subject of writing as well as the great theoretical and methodological breadth and diversity of the questions associated with it make it its very *own* field; *grapholinguistics*, thus, somewhat inadequately and unfairly ties it (and reduces it) to linguistics when not all writing-related aspects studied are actually linguistic in nature.<sup>32</sup> In other words, this line of criticism denounces the field's incorporation into (or appropriation by) linguistics that is terminologically insinuated by *grapholinguistics*.<sup>33</sup> However, in direct response to this, it can be argued that while the subject of writing is indeed multifaceted and can only be captured by a mixture of disciplines and associated methods, writing is, at its core, a linguistic phenomenon, i.e., the graphic manifestation of language<sup>34</sup>—which is not to say that it is not also a lot more than that. Against this background, the terminological focus on linguistics would be warranted even for an interdisciplinary *grapholinguistics*. Following this line of argument, it could also be claimed that while several aspects of writing can be studied without a consideration of its linguistic facets, a truly systematic—and arguably part of a comprehensive—analysis and theory of writing can only be achieved on the basis of a solid linguistic foundation. This is highlighted by linguist Elisabeth Stark (2022: 28) in her discussion of disciplinary limits and their relation to interdisciplinarity:

---

31. See also Joyce (2023: 140): “Meletis [...] suggests [...] that this designation has parallels with other subdisciplines of linguistics, such as sociolinguistics and psycholinguistics. While there is some merit in that observation, in contrast to the more interdisciplinary natures of both sociolinguistics and psycholinguistics, debatably, the term grapholinguistics fails to fully accord the study of writing with the central status that it deserves alongside the study of speech.” Cf. also Barbarić (2023: 119).

32. The fact that in the absence of institutionalization, grapholinguistics—or more generally the study of writing—does require some sort of ‘home’ discipline (or multiple such disciplines) to organizationally align with is discussed in Section 2.7.

33. In this context, Daniel Harbour (pers. comm., Oct. 2023) hypothesizes that having trained in formal linguistics could lead to finding the term less appealing: “In eschewing a name based on *linguistics*, we signal that we are stepping outside the linguistics in which we trained.”

34. Cf. also Meletis (2020: 8, emphasis in original): “[...] writing, following a narrow definition, refers only to those graphic (i.e., visual and/or tactile) ‘marks’ that represent language. This excludes marks that refer (directly) to ideas or extralinguistic referents. Writing is always intimately tied to language, and language is the subject of linguistics. The term *grapholinguistics* highlights this linguistic basis.”



Schrift als eigene Manifestationsform des Sprachlichen hat erst in jüngerer Zeit das systematische Interesse der Linguistik auf sich gezogen [...], und während die Beschreibung von Schriftsystemen und ihre Entstehung ebenso wie ihre gesellschaftliche und ökonomische Relevanz auch HistorikerInnen und im weiteren Sinne KulturwissenschaftlerInnen leisten können, kann nur eine Sprachwissenschaftlerin diesen Aspekten ein theoretisches Kapitel zur sprachwissenschaftlich fundierten Reflexion und Modellierung des Verhältnisses von Gesprochenem und Geschriebenem voranstellen. Schriftgeschichte, Orthographie und Typographie erfordern weiterhin eher wenig systematisches Wissen über die grundlegende Struktur menschlicher Sprache(n), wohl aber die Graphematik.<sup>35</sup>

Ironically, what Stark criticizes in her paper titled 'Warum es nur eine Linguistik gibt: Keine Interdisziplinarität ohne starke Disziplinen' ('Why there is only one linguistics: No interdisciplinarity without strong disciplines') is precisely that many scholars operate within interdisciplinary 'subdisciplines' that require linguistics or other neighboring disciplines to have permeable boundaries, which according to her causes conflation and ultimately a weakening of the participating disciplines. In her view, true and successful interdisciplinarity can only be achieved when disciplines are strictly and narrowly defined. Although she does not mention it explicitly, it can be assumed that she rejects an interdisciplinary grapholinguistics, as only a narrowly defined graphematics—indeed commonly conceived of as a (if not *the*) central subfield of grapholinguistics—is a truly linguistic matter.

Despite a focus on linguistic questions, it is precisely such an interdisciplinary interpretation of grapholinguistics that has been—at least in the German-speaking community—widely accepted, not least because of Dürscheid's textbook in which a chapter on graphematics is accompanied by chapters on, e.g., the history of writing, orthography, and typography—the topics Stark singles out as (predominantly?) non-linguistic. In other words, despite its terminological focus on linguistics, *grapholinguistics* denotes a field that is truly interested in all aspects of the linguistic phenomenon of writing—even if they are themselves non-linguistic. Thus, in recent publications, definitions such as the following can be found: "Schriftlinguistik (also known as grapholinguistics), a young linguistic subdiscipline that deals with the scientific study

---

35. "Writing as a separate form of manifestation of language has only recently attracted the systematic interest of linguistics [...], and while the description of writing systems and their emergence as well as their social and economic relevance can also be carried out by historians and, in a broader sense, cultural scholars, only a linguist can preface these aspects with a theoretical chapter on linguistically grounded reflection and modeling of the relationship between the spoken and the written. The history of writing, orthography, and typography still require rather little systematic knowledge of the basic structure of human language(s), but graphematics does" (my translation).

of *all* aspects of writing” (Condorelli 2022: 113, my emphasis). Further compelling evidence for the inclusivity of grapholinguistics is given by (socio)linguist Jürgen Spitzmüller who, starting with the third edition of Dürscheid’s *Einführung in die Schriftlinguistik* (published in 2006), contributes to the textbook a chapter covering typography. Now, many typographic aspects are not linguistic in a narrow sense, but this does not mean they lack communicative functions—quite to the contrary. Although the materiality of writing had long been dismissed by linguistics proper, it *is* studied in grapholinguistics, which is “die Teildisziplin, die es sich zur Aufgabe gemacht hat, eine umfassende theoretische Beschreibung schriftlicher Kommunikation zu leisten”<sup>36</sup> (Spitzmüller in Dürscheid 2016: 241). What definitely could still be debated (see the discussion of *philography* below) is whether grapholinguistics is also interested in aspects of writing that are non-communicative, which are logically also included when speaking of *all* aspects of writing. However, this discussion would in turn first necessitate answering the question of what such aspects may be—and whether, possibly, all aspects of writing are in fact in some way (not always, but in given contexts) communicatively relevant even when this is of course not always the perspective that is of primary interest.

A further challenge faced by the term *grapholinguistics*—and others with a similar history—shall be mentioned here: its above-described origin in Germanophone research, and thus its perceived boundedness to the German scholarly tradition,<sup>37</sup> which is likely part of the reason it is not (yet) found in many English-speaking publications (cf. Barbarić 2023: 123f.). It may be extreme and provocative to claim this, but it appears terms that do not originate in Angloamerican research traditions sometimes have a harder time being accepted by ‘originally’ Anglophone scholars. In some cases, if a term is not yet established in English-language research, scholars may even be oblivious of its existence. An illuminating example of this (cf. also Meletis in press) is an entry in the well-known linguistic blog *Language Log*. For context, it should be mentioned that in 2018, French mathematician, typographer, and linguist Yannis Haralambous initiated the conference series *Grapholinguistics in the 21st Century* (abbreviated as *G21C* and also known as */guafematik/*, see above) and later started the book series *Grapholinguistics and Its Applications* at Fluxus Editions, a publishing house he also founded. It is a mention of the second iteration of the *G21C* conference (held in 2020) that prompted Mark Liberman to publish a post titled ‘Grapholinguistics’ (originally

36. “[...] the sub-discipline that has set itself the task of providing a comprehensive theoretical description of written communication” (my translation).

37. In this context, the pioneer status of German-language grapholinguistic research is occasionally mentioned (cf. Meletis 2020; Neef 2021; Meletis & Dürscheid 2022).

in double quotation marks, which serve a distancing function here) in the *Language Log*. In it, Liberman first cites a passage from the conference announcement in which Haralambous comments on grapholinguistics' little-known status:

*Grapholinguistics* is the discipline dealing with the study of the written modality of language. At this point, the reader may ask some very pertinent questions: 'Why have I never heard of grapholinguistics?' 'If this is a subfield of linguistics, like psycholinguistics or sociolinguistics, why isn't it taught in Universities?' 'And why libraries do not abound of books [sic] about it?'

After giving this quote, Liberman proceeds to answer the first of these questions: "Speaking for myself, I'll answer: We've never heard of grapholinguistics because you just made up the word." He goes on to remark that "[u]nder headings like 'Writing Systems', the issues involved are widely taught in universities," likely implying that there is no need for the term *grapholinguistics*. Also, he lists a number of—exclusively English-language—monographic works on writing systems and contends that "there have been plenty of previous objections to the treatment of writing systems as entirely secondary, derivative, and even negligible," citing a lengthy passage from Nunberg's (1990) *The linguistics of punctuation*. Finally, he writes, "[s]o I guess that at G21C 2020 we'll learn that everything old is new again..." insinuating that grapholinguistics as a discipline attempts to reinvent the wheel and is not critically aware of and founded on important works in the study of writing—even if these had of course not been published under the heading of *grapholinguistics*. In a footnote, Liberman lists works in which the term occurs that he found on Google Scholar, including Sariti (1967) or, in a very different sense, Platt (1974), but oddly fails to mention Neef (2015) or Meletis (2018), articles published before 2020 that carry the term in their titles and are shown (at the time of the writing of this article) on the first result page for 'grapholinguistics' in Google Scholar.

In conclusion, what Liberman's blog post proves is not that Haralambous has made up the word or the field associated with it, but that—highlighted also by numerous comments made by users under the post—researchers in English-language research communities may be oblivious to its existence and rich history. To close with a more hopeful counterexample, however, it is worth mentioning that in 2020, the term was adopted by Australian linguists Piers Kelly and Arvind Iyengar, who, in the abstract of their conference talk 'What is writing? Grapholinguistics as a field of scholarly inquiry' not only acknowledge that writing is an up-and-coming subject in linguistics, archaeology, and anthropology, but also associate the resurgence of interest in writing with the 'new'

term *grapholinguistics*: “This is affirmed by the recent acceptance of a new name for the study of writing systems: grapholinguistics.”<sup>38</sup>

## 2.6. Script(ur)ology: A New Term for a New Field?

One of the shorter sections of this paper shall be devoted to a candidate designation that was coined rather recently in the context of French semiotics (or semiology): *script(ur)ology*. In the relevant sense presented here, it was introduced in a special issue of the French journal *Signata: Annales des sémiotiques / Annals of Semiotics* entitled ‘Signatures. (Essays in) Semiotics of Writing’ edited by Jean-Marie Klinkenberg and Stéphane Polis. In the issue’s introduction,<sup>39</sup> they write:

Writing is envisioned here in its generality, as a semiotic system that mediates between the linguistic and spatio-iconic realms. Indeed, based on detailed analyses of the semiotic functions fulfilled by graphemes, the aim of this issue is admittedly to identify criteria and principles that could be used for developing a typology of writing. As such, the volume ambitions to contribute to a ‘general scriptology’, a discipline already explored by pioneering works, such as the ones by Roy Harris or Anne-Marie Christin, to name but a few of the directions that this endeavor might pursue.

Conceptually, the envisioned scriptology<sup>40</sup> is, due to its semiotic conception, broader than, for instance, a linguistic graphem(at)ics, since it is—as Klinkenberg and Polis explicitly mention—certainly also interested in spatial and iconic aspects of the written modality that are usually neglected by graphem(at)ics (and/or relegated to neighboring sub-disciplines such as graphetics, cf. Meletis 2015). However, at the same time, scriptology may be more narrowly conceived than grapholinguistics, as usage-based and communicatively relevant aspects such as sociolinguistic or psycholinguistic ones remain unmentioned.<sup>41</sup> Terminologically, as a Latin-Greek hybrid (which in its linguistic composition is the mirror image of Greek-Latin *grapholinguistics*), *scriptology* relies

---

38. See <https://rune.une.edu.au/web/handle/1959.11/30186> (accessed October 30, 2022).

39. Alas, the PDF or print version of the introduction was not available to me, only the online version (<https://journals.openedition.org/signata/1274>, accessed October 31, 2022), which is why this passage is cited without page numbers.

40. Condorelli (2022: 116, emphasis in original) also mentions a different meaning of (French) *scriptologie*: “Generally speaking, *scriptologie* has been used as a framework of inquiry for studying the Gallo-Romance and Italo-Romance dialectal areas and, although less comprehensively, the Ibero-Romance area.”

41. Note, however, that Condorelli (2022: 115), for example, still interprets the two as more or less synonymous: “[...] *scriptology*, which [...] corresponds to the general area of writing theory that contemporary linguists call grapholinguistics.”

on the polysemous term *script* that is associated with many a concept in linguistics and beyond (see also above) and elevates it to an entire 'study' of writing by using the suffix *-logy*. What the authors mean by *script* remains—at least in their introduction—implicit, although several passages such as the following allow drawing conclusions: "The traditional descriptions of writing systems—which classify scripts in broad categories (alphabets, Abjads [sic], syllabic scripts, logographic scripts, etc.)—necessarily simplify their richness and intrinsic hybridity." Here, as is so often the case in literature on writing, the terms (and associated concepts) *writing system* and *script* occur in close proximity and are likely conflated by being used more or less synonymously, here with the meaning 'type of writing system', examples of which are the listed categories alphabet, abjad, etc. In my reading, biased by my own theoretical conception of writing, *writing system* denotes the system of writing in/for a specific language (such as English), while I interpret *script* in a material sense as a historically developed set of basic shapes (such as Roman or Cyrillic script) that can theoretically be coupled with any language.<sup>42</sup>

Confusingly, the authors' introduction of new terminology is compromised by an unexplainable case of inconsistency when—in inconformity with the issue's introduction—in their following 'texte intégral' in which they sketch their envisioned field, the central term suddenly reappears one syllable richer—as *scripturology*. In its French original, this main article is titled 'De la scripturologie' (an homage to Derrida?), while the English translation<sup>43</sup> is given as 'On scripturology'. In the latter, Klinkenberg & Polis provide these definitions for the newly christened field:

In this contribution we present the principles and parameters of a discipline which remains—in our intended meaning—largely yet to be established: *scripturology*. This discipline concerns the study of different facets of writing, perceived in its generality, as the semiotic apparatus articulating language facts and spatial facts. (Klinkenberg & Polis 2018: 57, emphasis in original)

Scripturology is understood as a general theory targeting the establishment of a semiotic typology of writing systems. Its horizon is therefore compatible, within the study of writing, to that of linguistic typology. (Klinkenberg & Polis 2018: 58)

These passages reveals that they consider scripturology to be part of a larger study of writing, confirming the above assessment that it

---

42. Cf. also Coulmas (1996b: 1380, emphasis in original): "Script refers to the actual shapes by which a writing system is visually instantiated. [...] Every writing needs for its materialization a script, but there is no necessary link between a particular script and a particular writing system". But see Gnanadesikan (2017) for a use in line with Klinkenberg's and Polis'.

43. The English translation was prepared by Todd J. Gillen.

is defined more narrowly than grapholinguistics. Terminologically, although only separated by one syllable, *scripturology* differs quite significantly from *scriptology* as it is not tied to *script* but rather to a different word, as the authors explicitly note:

The term retained for designating this domain of study is a blended compound, forged from the Latin deverbal noun *scriptura* (which refers both to the ‘written thing’ and to the ‘composition’) and from the Greek suffix *-logie* (which performatively establishes the scientific character of the field); this designation indexes, in some way, the hybrid and heterogeneous character of the domain of study that we bring together and unify under this banner. (Klinkenberg & Polis 2018: 58, emphasis in original)

At the specific level, one could ask the question of whether (and why) a new term is needed for what essentially appears to be a semiotically broader approach to writing system typology.<sup>44</sup> More globally, what can be discussed in this context is the general decision to coin a new term. Arguably, proposing a new designation for a field is intended to echo the novelty of one’s idea; as Klinkenberg & Polis (2018: 57) emphasize, in their meaning, the discipline has “yet to be established”. Tying a new name to it—not unlike Christa Dürscheid did with her *Einführung in die Schriftlinguistik*, although *Schriftlinguistik* was not entirely new but rather unestablished—is, on the one hand, meant to contribute to the establishment of the field. On the other hand, we find another motivation rooted in the sociology of science (or rather, at a meta-level, academia): coining a new label—or successfully reappropriating it, see Derrida and ‘his’ *grammatology*—has the potential to tie the founder to the named discipline in quite a profound way. This can go awry when the term is not adopted by others and buried in oblivion; if, however, it is accepted and comes into widespread use, it can, by association, automatically cement the coiner’s status as an authority in the field.

To close this section, as was done in the preceding ones, different, potentially even non-writing related uses of the discussed term shall be mentioned briefly. In the case of Latin *scriptura*, of course, it is rather obvious which other meaning—besides ‘something written’—is a candidate for interference, as it has prevailed as the meaning of modern English *scripture*. Indeed, *scripturology* can be found—albeit admittedly not often—in this theological reading, an example being Mohsen Goudarzi Taghanaki’s PhD thesis *The second coming of the book: Rethinking Qur’anic scripturology and prophethology*, in which *scripturology* is defined as “a new interpre-

---

44. In Joyce & Meletis (2021), ‘traditional’ writing system typology’s narrow focus on the linguistic levels that written units relate to (yielding categories such as phonography and morphography) is likewise criticized as being simplistic and reductive, and alternative criteria for other types of (also psycholinguistic and sociolinguistic) typologies are proposed (cf. also Meletis 2021b)—however, no new term is introduced.

tation of the Qur'an's conception of scriptural [...] history" (Goudarzi Taghanaki 2018: iii). A related definition is also provided in Tan (1982: 51): "Scripturology is a rather generic designation of the study of all written bases or scriptures or religions such as the Bible for Christianity, the Koran for Islam, the Tend-Avesta for Zoroastrianism, the Vedas for Hinduism, the Tripitaka for Buddhism, the Kojiki or Nihonji, for Shintoism, and others."

## 2.7. Philography: An (Old) New Term and the Future of an Identity Crisis

In recent years, as the study of writing is gaining traction and a more international community is forming—thanks to conference series such as the *Association for Written Language and Literacy* workshops and others—the field's designation has become the target of renewed debate. Especially the recent—prominent and highly visible—adoption of *grapholinguistics* in the title of the *Grapholinguistics in the 21st Century* conference series and the associated impression that it is in the process of winning this terminological battle have resulted in both an actual increase of occurrences of the term and the fact that it is more vocally scrutinized. The latter also stimulates the (renewed) discussion of alternative terms in which the present paper can be contextualized and that also at times produces new proposals. At this point, then, the non-exhaustive treatment of different candidates shall be closed with the presentation of such a 'new' (if in fact pre-existing) term that has been suggested in this context: *philography*.

In informal chats during conference breaks (at the 12th workshop of the Association for Written Language and Literacy in Cambridge in 2019), Amalia Gnanadesikan and Daniel Harbour expressed their reservations about *grapholinguistics* and brainstormed possible alternatives, agreeing on *philography* as a suited candidate.<sup>45</sup> When invited to elaborate on their preference, they explained as follows:

I do like 'philography', though. I like the appeal to precedent in 'philosophy' and 'philology'. While I have no objection to the use of 'grapholinguistics' when it is applicable [...], I like the fact that 'philography' focuses on writing in and for itself, not just when it is a subfield of linguistics. Thus I see it as a wider word than 'grapholinguistics'. It is both more inclusive (not just linguistics) and more focused on its actual subject (writing itself in all its aspects) [...]. (Amalia Gnanadesikan pers. comm., Oct. 2022)

'Philography' suggests a study that crosscuts these disciplines [neurolinguistics, psycholinguistics, sociolinguistics, historical linguistics, ..., DM], just as philosophy and philology do. And, like philosophy especially, it can

---

45. Harbour has already used the term—in the form of the adjective *philographic*—in Harbour (2021: 201).

bleed at the edges. Just as there is philosophy of art, so there is artistic use of scripts, of typography. I want these areas to be included in our discipline and I see ‘philography’ as opening that door in a way that ‘grapholinguistics’ doesn’t. (Daniel Harbour pers. comm., Oct. 2022)

Again, the terminological side of the story is—evidently and justifiably—tied to the (self-)conception of the field and the delimitation of its scope. Formally, the all-Greek *philography* is comprised of *philo-*, from Ancient Greek φίλος ‘loving, beloved, dear’, and *-graphy*, which as the obvious writing-related component occurs here at the end of the term (for a change). ‘Love of writing’ as the meaning of this undeniably elegant<sup>46</sup> term is indeed fitting to denote a field that deals with all aspects of writing. And, as Gnanadesikan and Harbour point out, it is comprehensive, i.e., inclusive of all possible facets of writing and the perspectives and methods studying them, which in this respect makes it superior to the (at least terminologically) linguistics-focused *grapholinguistics*.<sup>47</sup> Or does it?

This is an appropriate point to dwell on this question of inclusivity, which in its complexity surpasses the mere choice of a label for the field. Indeed, while all the many disciplines and scholars working on matters of writing should be welcomed by ‘the’ study of writing, what can be observed with almost all attempts at coining a designation outlined in this paper is that they usually still originate in an existing and established discipline—and in most cases, this is linguistics. A truly inclusive and balanced philography, by contrast, would favor no discipline participating in it, which, pessimistically, could lead to a rather fragmented state of the field with a weak common thread or shared core. If all perspectives on writing are valid, what is the main one? Does there need to be one? In theory, and when it comes to the actual study of the subject of writing, no. However, this question is not only of theoretical nature but one with paramount practical, e.g., institutional implications that could prove decisive when considering the future of the field (cf. also Meletis 2021a). Put simply: Where would philography fit in? This question is justified as we are possibly too late in the game (if there’s ever such a point) to aim to shape an entirely new field that we eventually—and

---

46. It would be naive to think that aesthetic considerations do not also feature prominently in terminological discussions. This is underlined by Harbour’s (pers. comm., Oct. 2022) personal assessment that he finds *grapholinguistics* unappealing.

47. Another aspect that Harbour (pers. comm., Oct. 2023) mentions is the naming of potential subdisciplines: “Philography can and should have subdisciplines, such as the neurolinguistics, psycholinguistics, and sociolinguistics of script use. It is perfectly natural for me to refer to these specialisms as *neuropsycholinguistics*, *psychophilography*, and *sociophilography*, just as it is to talk about *neurolinguistics*, *psychosemantics*, and *sociophonetics*, all established terms in the field. Parallel names based on *grapholinguistics* are plain awful. Fields called *neurographolinguistics*, *psychographolinguistics*, or *sociographolinguistics* deserve to fall stillborn from the press [...]”



rather sooner than later—expect to translate into chairs and journals and conferences and everything else associated with established fields. What the study of writing *is*—that's not just a question asked by (and from within) a field that has an ongoing identity crisis but likewise a question of where the field should be 'put' organizationally, also concerning where it has the best chances to thrive. If viewed from a different perspective, it's also a question of 'ownership': *grapholinguistics* insinuates that linguistics has a prerogative with respect to the study of writing. At the other end of the spectrum, *philography*—at least terminologically—makes it a disciplinary orphan. Of course, this discussion is a lot more complex than sketched here, and an inclusive philography can certainly have specific focuses and/or can be institutionally connected to an established discipline.

Interestingly, as foreshadowed above, *philography* is not a completely new term,<sup>48</sup> and it has occasionally been mentioned in discussions of a name for the study of writing, for example by Gelb (see Section 2.1). Specific uses appear to be rare, however. One such occurrence of the term—in which it is not straightforwardly defined—is found in Andreas Gottschling's (1881/1882) 'Über die Philographie' ('On philography').

Finally, another meaning of *philography* that its use as a designation for the study of writing must compete with is "the collecting of autographs, esp. those of famous persons".<sup>49</sup>

### 3. The Common Threads

In this section, several common threads characterizing naming processes in the study of writing will be presented in the form of a critical summary. Note that these are not mutually exclusive but overlap and interact in complex ways, with their separation here only serving as an idealization for illustrative purposes.

(1) Firstly, what we commonly find is mentions of the novelty or unestablished nature (and/or marginal status) of the field that is to be named: When Gelb (1952) proposed *grammatology* and initially even included it in the title of his book,<sup>50</sup> there certainly had already existed research on writing in various forms. However, with the fittingly named *A study of writing*, as is probably unanimously accepted among scholars

---

48. There is even a dedicated Wiktionary entry: <https://en.wiktionary.org/wiki/philography> (accessed November 1, 2022).

49. Cf. <https://www.collinsdictionary.com/dictionary/english/philography> (accessed November 1, 2022).

50. Interestingly, the subtitle *The foundations of grammatology* was dropped from the second edition (1963).

of writing, he ushered in a new era in which research on writing became more focused and more about writing in and of itself. From that point on, *grammatology* was the designation to beat—until Derrida’s famous borrowing of it in the 1960’s, that is. The reason it did not prevail pre-Derrida is, however, most likely not of terminological nature but rather due to the marginal status writing had as a subject in linguistics. In other disciplines, ironically, the situation was the polar opposite: Specific philological branches with rich research traditions, especially ones with a focus on historical languages (among them the archaeology- and anthropology-infused assyriology that Gelb was invested in), were sometimes so focused on written documents, written language, and writing in general that coining a separate term for its study likely appeared superfluous and counterintuitive. Against this background, it is unsurprising that most attempts at coining a term discussed in this paper can be contextualized within linguistics, because there, the study of writing actually needed to be emancipated and had to prove itself. Ultimately, however, writing remained a linguistic niche topic for so long that the novelty of the field or different approaches in it kept being underlined. In 2002, Dürscheid mentioned that *Schriftlinguistik* had not yet entered the canon of linguistic terminology, and in 2018, Klinkenberg & Polis (2018: 57) named a discipline that “in [their] intended meaning” is “largely yet to be established” *script(ur)ology*. This underlines an important function ascribed to the naming process: It is intended to have a constitutive force. A field that has no fixed and accepted name may be unestablished for this precise reason, so performatively giving it a name is meant to provide it with a more stable identity (cf. also the examples in Powell et al. 2007). Therefore, and given the still ongoing debate about the field’s name, the question can and should be asked of what this tells us about the state of the field.

One more aspect that should be mentioned here as it is closely related to pointing out the unestablished nature of the field is that—as was discussed in the context of *script(ur)ology*—coining a new designation is also a process of claiming it as one’s own. This can be seen at the disciplinary level, when scholars want to claim the field for their discipline or at least highlight the prominence or priority of their discipline in studying writing (cf. *grapholinguistics*), but also at the individual level, when specific scholars want to be seen as the ones who elevate the field or a specific approach to a more established status (cf. *script(ur)ology*).

(2) In the context of presenting their term of choice, authors often also list the existing alternatives and take this opportunity to point out their shortcomings. The ubiquity of this practice is not accidental but rather systematic as it is a symptom of the awareness surrounding this central terminological question plaguing the study of writing. By doing this, authors also strengthen further the hierarchies created by arguing

for their term of choice, as downplaying the suitability of possible other candidates serves to highlight the inevitability of their candidate.

(3) The coining of designations for entire fields appears to follow certain principles, one of them mandating the designation be as semantically fitting and transparent as possible. Also, it should fit in with existing designations for other (established) fields. The former principle is the reason for the recurrence of *graph-* in various forms (both as *grapho-* and *-graphy*) and positions (both in initial and final positions). The latter principle, on the other hand, straightforwardly explains the use of productive bound morphemes such as *-logy* or *-nomy*. Problems with the perceived suitability of names for the study of writing, now, arise precisely when these principles are not adhered to: *grammatology* has been criticized because of the narrower meaning that *gramma-* can have ('letter'), let alone its possible association with *grammar* and the writing of grammars (cf. the reading of *grammatology* in Zaefferer 2006; note that in this meaning, it can also be found as *grammaticography*). Similar reasons have been stated for the unsuitability of *scriptology* and *script linguistics*, as *script* has many definitions which are in most cases also narrower than that of *writing* in general, and the former's alternative variant *scripturology* evokes the wrong association. Conversely, *grapholinguistics* narrows it down at the other end as switching the neutral *-logy* or *-nomy* for the name of a specific discipline leads to a whole slate of problems (see also (6)).

The described principles are not confined to naming processes for/in the study of writing. Thus, the field has no monopoly over elements such as *graph-* and *-logy*, which of course is the reason we find so many of the terms presented here used in different contexts and with distinct meanings. Some of these meanings, such as the ones of *graphology* and arguably also *grammatology*, had either previously been dominant (as in the case of *graphology*) or have prevailed over time (*grammatology*).

(4) Another terminological issue in the narrow sense is the ambiguity typical of many terms in linguistics (and other disciplines): the phenomenon and the field/branch/discipline studying it are referred to by the same name, which applies to the most established of designations such as *phonology*, *morphology*, and *syntax*. *Grammatology*, *graphonomy*, *graphology*, *graphem(at)ics*, *orthography*—all of these terms can denote phenomena of writing, in most of such uses something along the lines of 'the written level of language' or 'the graphic component of language', as well as the subbranches studying this very level/component. Notably, this latter meaning is sometimes expanded as the terms can also be used more broadly: *graphem(at)ics*, then, can encompass more than the study of the graphem(at)ic module of language. This is rather seldom the case, and all of the mentioned terms are commonly and predominantly associated with language and linguistics, insinuating that the study of writing is only concerned with its linguistic aspects.

That being said, with respect to broader alternatives, *grapholinguistics* is simultaneously wider in its meaning—according to most definitions, it is supposed to study all aspects of writing, not only writing as a component of language—and as narrow as (or even narrower than) these terms, as it is directly and visibly bound to linguistics, lending it a restricted and exclusivist aftertaste (cf. (6)). And *writing systems research*, as has been argued above, may appear broad but has its own drawbacks, as ‘writing system’ is likewise connoted linguistically and excludes aspects that could intuitively be judged as ‘non-systematic’ from a descriptive linguistic perspective.

(5) A challenge mentioned in the context of *Schriftlinguistik* and its slow and bumpy transition into a scrutinized English-language equivalent is the hold that Anglophone research communities seem to have over terminology. This has arguably not always been the case as English has only gradually advanced to an academic lingua franca, a process that has led to questionable and problematic maxims such as ‘if you want to be read (internationally), you need to publish in English’, an issue that appears even more exacerbated with respect to terminology. Against this background, terms that were introduced in other languages and, likely more importantly, whose introduction and adoption were embedded in a non-Anglophone research culture and tradition, are possibly at a particular disadvantage. In the case of terms for the study of writing originating in other cultures, not only must a fitting English translation be found that is accepted by scholars who want to participate, but research that has previously been carried out under this banner often continues to be (made) invisible.

A failure to look beyond one’s horizon or outside of one’s language may result in the complete oblivion of a possible designation. For *grapholinguistics*, this was shown with a blog post by an American scholar claiming that the word had just been made up (see Section 2.5). I want to mention another illuminating example that is, however, not located (solely) at the terminological but—which appears even more severe—at the conceptual level: In 1991, Peter T. Daniels published a paper titled ‘Is a structural graphemics possible?’, ultimately concluding that there cannot be such a field and thus negating his question; in 1994, he received a reply by Earl M. Herrick, who also devoted much of his research to questions of writing and gave his rebuttal the title ‘Of course a structural graphemics is possible!’. As I tried to show elsewhere (Meletis in press), their entire discussion about the possibility of a structuralist approach to writing—while certainly raising valid and to this day crucial points about the field—seems weirdly anachronistic for scholars socialized in a German(ist) linguistic tradition since at the beginning of the 1990s, questions of graphem(at)ics had long been intensively discussed and partially even settled in the German grapholinguistic community. I named the article in which I present and historiographically contex-

tualize their dispute 'There had already been a structural graphemics', which basically says it all. Ergo: cultural and linguistic boundaries are real, and they can pose major challenges in the establishment of fields and terminology (cf. Meletis 2021a).

Sometimes, meanings also get lost in translation, impeding the cross-linguistic applicability of certain terms. The above-mentioned *orthography*, for instance, has a broader and more descriptive meaning in Anglophone literate cultures than it does in German. A designation such as *historical orthography* will, thus, not as easily be accepted by scholars rooted in Germanophone traditions, who in this case indeed prefer *historical graphematics*. This also shows that the dismissal of terms can also work in the other direction, although the involved hierarchical dynamics in the two scenarios are certainly not equal.

(6) A central issue that seems to be taken for granted for the study of writing and is debated with respect to a suitable designation is—in more than one respect—inclusivity. As mentioned in (4), many of the terms collected in this paper are—for one reason or the other—tied to a specific discipline: linguistics. This applies to the maximum degree to *grapholinguistics* although the associated field—as evidenced also by the interdisciplinary conference series *G21C*—actually prides itself on including all disciplines along with their research questions and the theories and methods employed in approaching these. Yet, it is understandable and certainly valid that a psychologist working primarily on visual aspects of reading or an art historian researching the appearance of writing in different types of art would refuse to describe their work as 'grapholinguistic'.<sup>51</sup> Against this background, it may be striking but ultimately unsurprising that inclusive definitions of grapholinguistics and attempts to motivate others to adopt it stem almost exclusively from linguists.

Debates surrounding terminology reflect negotiations of power and ownership, which means that from this perspective, a label as neutral as possible would be a preferable democratic choice. *Philography* has been named as one possibility for such a neutral designation. However, the question that was raised in this context was whether the adoption of such a neutral term would actually avert negotiations of power within the study of writing. If the field is not to be seen as a fragmentary collection of those subfields of linguistics, psychology, anthropology, etc. that deal with writing but an independent field that incorporates all of

---

51. In this context, an aspect that was altogether omitted in the present paper shall at least be mentioned: the corresponding terms that stand for people. In Meletis (2021a), for example, I call myself a 'grapholinguist', and at least 'grammatologist' and 'philographer' are also imaginable (with 'grammatologist' actually sporadically being used in the literature). These terms are even more contentious as they are tied not only to disciplines but to specific people and their individual self-conception as scholars.

those into a bigger and coherent picture, then adopting a neutral term truly requires (re)shaping the field's identity around said picture. This is a complex process that implicates many more questions such as: Do all disciplines even want to 'sit at the table' (and to an equal degree)? What is the definition of 'writing' that such a study of writing in which all disciplines are truly equal relies on? In which department(s) would such a discipline have its home, or do we really aspire independence to such a degree that it would need its own new department? More practically: Who would organize and fund conferences? In philosophical thought experiments like this, no questions are disallowed. In reality, however, when it comes to an actual implementation, most scholars of writing likely see no point in pursuing (likely risk-laden) answers to them.

#### 4. The Future

In order to give an outlook, we need to first sum up where we stand right now: As of yet, there is no widely accepted designation for the study of writing. As this paper attempted to show, this is certainly not due to a shortage of possible candidates. For each of them, however, compelling reasons speaking against a more widespread and uniform adoption can be found. Interestingly, all discussed terms still live on as each pops up sporadically in the literature, referring to the study of writing in—sometimes unexpected—contexts, at times explicitly linking to an existing terminological tradition, at others simply being coined due to terminological obviousness. Indeed, given that most of them are rather transparent and thus justifiable compositions, their continued (co-)existence is rather unsurprising and will likely continue. In general, the terminological discussion surrounding the study of writing as captured in this paper is a positive reflection of the resilience of both the field and continuous attempts at further establishing it. This does raise the question of whether we are stuck in an unproductive loop of recycling terms and arguing for their suitability, though. Conferences help in slowly forming an international community out of many diverse communities, and in this context, the name of the field is indeed only one—and not the most important—issue that needs settling. Other questions—theoretical, methodological, ones regarding the politics of science and academia—must likewise be faced, and it is unpredictable how they will influence terminology... and vice versa. Ultimately, an unambiguous, inclusive designation that pleases everyone may be a desideratum or wishful thinking. That's because with respect to scientific terminology, the answer to 'What's in a name?' is clearly: a lot.

## References

- Allén, Sture (1971). *Introduktion i grafonomi. Det lingvistiska skriftstudiet*. Stockholm: Almqvist & Wiksell.
- Althaus, Hans Peter (1980). "Graphemik." In: *Lexion der Germanistischen Linguistik*. Ed. by Hans Peter Althaus, Helmut Henne, and Herbert Ernst Wiegand. Tübingen: Max Niemeyer, pp. 142–151.
- Ashcroft, W. D. (1989). "Constitutive graphonomy. A post-colonial theory of literary writing." In: *Kunapipi* 11.1, pp. 58–73.
- Augst, Gerhard, ed. (1985). *Graphematik und Orthographie. Neuere Forschungen der Linguistik, Psychologie und Didaktik in der Bundesrepublik Deutschland*. Vol. 2. Theorie und Vermittlung der Sprache. Frankfurt a. M., Bern, New York: Peter Lang.
- ed. (1986). *New trends in graphemics and orthography*. Boston, Berlin: De Gruyter.
- Ballhorn, Friedrich (1947). *Alphabete orientalischer und occidentalischer Sprachen. Zum Gebrauch für Schriftsetzer und Correctoren*. Leipzig: Brockhaus.
- (1961). *Grammatography. A manual of reference to the alphabets of ancient and modern languages*. London: Truebner & Co.
- Barbarić, Vuk-Tadija (2023). "Grapholinguistics." In: *The Cambridge handbook of historical orthography*. Ed. by Marco Condorelli and Hanna Rutkowska. Cambridge: Cambridge University Press, pp. 118–137.
- Bazell, Charles E. (1956). "The Grapheme." In: *Litera* 3, pp. 43–46.
- Berg, Kristian (2019). *Die Graphematik der Morpheme im Deutschen und Englischen*. Berlin, Boston: De Gruyter.
- Berg, Kristian and Martin Evertz (2018). "Graphematik – die Beziehung zwischen Sprache und Schrift." In: *Linguistik – Eine Einführung (nicht nur) für Germanisten, Romanisten und Anglisten*. Ed. by Stefanie Dipper, Ralf Klabunde, and Wiltrud Mihatsch. Berlin: Springer, pp. 187–195.
- Börner, Wolfgang (1977). "Besprechung von Hellberg Stefan, Graphonomic rules in phonology. Studies in the expression component of Swedish." In: *Indogermanische Forschungen* 82, pp. 335–343.
- Condorelli, Marco, ed. (2020). *Advances in historical orthography, c. 1500–1800*. Cambridge: Cambridge University Press.
- (2022). *Introducing historical orthography*. Cambridge: Cambridge University Press.
- Coulmas, Florian, ed. (1996a). *The Blackwell encyclopedia of writing systems*. Oxford: Wiley-Blackwell.
- (1996b). "Typology of writing systems." In: *Schrift und Schriftlichkeit/Writing and its use*. Ed. by Hartmut Günther and Otto Ludwig. Vol. 10.2. Handbooks of Linguistics and Communication Science. Berlin, Boston: De Gruyter, pp. 1380–1387.
- Crystal, David (1980). *A first dictionary of linguistics and phonetics*. London: Deutsch.

- Crystal, David (1997 [1987]). *The Cambridge encyclopedia of language*. 2nd ed. Cambridge: Cambridge University Press.
- (2003). *A dictionary of linguistics and phonetics*. 5th ed. Oxford: Blackwell.
- Crystal, David and Derek Davy (1979 [1969]). *Investigating English style*. London: Longman.
- Daniels, Peter T. (1990). "Fundamentals of grammatology." In: *Journal of the American Oriental Society* 110.4, pp. 727–731.
- (1991). "Is a structural graphemics possible?" In: *LACUS Forum* 18, pp. 528–537.
- (1996). "The study of writing systems." In: *The world's writing systems*. Ed. by Peter T. Daniels and William Bright. Oxford: Oxford University Press, pp. 3–17.
- (2009). "Grammatology." In: *The Cambridge handbook of literacy*. Ed. by David R. Olson and Nancy Torrance. Cambridge: Cambridge University Press, pp. 25–45.
- (2018). *An exploration of writing*. Bristol: Equinox.
- (2021). "Foundations of graphonomy. A linguist addresses psychologists." In: *Journal of Cultural Cognitive Science* 5, pp. 113–123.
- Daniels, Peter T. and William Bright, eds. (1996). *The world's writing systems*. Oxford: Oxford University Press.
- De Chadarevian, Soraya (2002). *Designs for life. Molecular biology after World War II*. Cambridge: Cambridge University Press.
- DeFrancis, John (1950). *Nationalism and language reform in China*. Princeton: Princeton University Press.
- Derrida, Jacques (1967). *De la grammatologie*. Paris: Éd. de Minuit.
- (1997 [1976]). *Of grammatology*. Baltimore: Johns Hopkins University Press.
- Dürscheid, Christa (2016). *Einführung in die Schriftlinguistik*. 5th ed. Mit einem Kapitel zur Typographie von Jürgen Spitzmüller. Göttingen: Vandenhoeck & Ruprecht.
- Eckardt, Andre (1965). *Philosophie der Schrift*. Heidelberg: Julius Groos.
- Ehlich, Konrad (2001). "Graphemics/[Transindividual] graphology." In: *Loss of communication in the information age*. Ed. by Rudolf de Cillia, Jürgen Krumm, and Ruth Wodak. Vienna: Verlag der Österreichischen Akademie der Wissenschaften, pp. 61–76.
- (2007). *Sprache und sprachliches Handeln*. Berlin, Boston: De Gruyter.
- Elmentaler, Michael (2018). *Historische Graphematik des Deutschen. Eine Einführung*. Tübingen: Narr.
- Fleming, Juliet (2016). *Cultural graphology. Writing after Derrida*. Chicago: University of Chicago Press.
- Fuhrhop, Nanna and Jörg Peters (2013). *Einführung in die Phonologie und Graphematik*. Stuttgart: Metzler.



- Gallmann, Peter (1985). *Graphische Elemente der geschriebenen Sprache. Grundlagen für eine Reform der Orthographie*. Vol. 60. Reihe Germanistische Linguistik. Berlin, Boston: De Gruyter.
- Gelb, Ignace J. (1963 [1952]). *A study of writing*. 2nd ed. Chicago: University of Chicago Press.
- Glück, Helmut (1987). *Schrift und Schriftlichkeit. Eine sprach- und kulturwissenschaftliche Studie*. Stuttgart: Metzler.
- ed. (1993). *Metzler Lexikon Sprache*. Stuttgart: Metzler.
- Gnanadesikan, Amalia E. (2017). "Towards a typology of phonemic scripts." In: *Writing Systems Research* 9.1, pp. 14–35.
- Gómez-Jiménez, Eva (2015). "An introduction to graphology. Definition, theoretical background and levels of analysis." In: *miscelánea* 51, pp. 71–85.
- Gottschling, Andreas (1881/82). "Über die Philographie." In: *Zeitschrift für Orthographie, Orthoepie und Sprachphysiologie* 2, pp. 200–201.
- Goudarzi Taghanaki, Mohsen (2018). "The second coming of the book. Rethinking Qur'anic scripturology and prophetology." PhD thesis. Harvard University.
- Günther, Hartmut (1988). *Schriftliche Sprache. Strukturen geschriebener Wörter und ihre Verarbeitung beim Lesen*. Vol. 40. Konzepte der Sprach- und Literaturwissenschaft. Berlin, Boston: De Gruyter.
- Günther, Hartmut and Otto Ludwig, eds. (1994/96). *Schrift und Schriftlichkeit/Writing and its use*. Vol. 10.1 and 10.2. Handbooks of Linguistics and Communication Science. Berlin, Boston: De Gruyter.
- Hall, Robert A. (1960). "A theory of graphemics." In: *Acta Linguistica* 8, pp. 13–20.
- Hamp, Eric P. (1959). "Graphemics and paragraphemics." In: *Studies in Linguistics* 14.1, pp. 1–5.
- Haralambous, Yannis (2019). "Approaches to and applications of graphemics." In: *Methods and interdisciplinarity*. Ed. by Roger Waldeck. London, Hoboken: ISTE, John Wiley & Sons, pp. 149–169.
- Harbour, Daniel (2021). "Grammar drives writing system evolution. Lessons from the birth of vowels." In: *Grapholinguistics in the 21st Century—2020, Proceedings*. Ed. by Yannis Haralambous. Vol. 4. Grapholinguistics and Its Applications. Brest: Fluxus Editions, pp. 202–221.
- Hasse, Johann Gotfried (1792). *Versuch einer griechischen und lateinischen Grammatologie*. Königsberg: Friedrich Nicolovius.
- Hellberg, Staffan (1974). *Graphonomic rules in phonology. Studies in the expression component of Swedish*. Vol. 7. Nordistica Gothoburgensia. Göteborg: Acta Universitatis Gothoburgensis.
- Herrick, Earl M. (1994). "Of course a structural graphemics is possible!" In: *LACUS Forum* 21, pp. 413–424.
- Hjelmslev, Louis (1947). "Structural analysis of language." In: *Studia Linguistica* 1.1, pp. 69–78.

- Hockett, Charles F. (1951). "Review of 'Nationalism and language reform in China,' by J. DeFrancis." In: *Language* 27.3, pp. 439–445.
- (1952). "Speech and writing." In: *Report of the third annual round table meeting on linguistics and language teaching*. Ed. by Salvatore J. Castiglione. Vol. 2. Monograph Series on Language and Linguistics. Washington: Georgetown University Press, pp. 67–76.
- (2003). "Two lectures on writing." In: *Written Language & Literacy* 6.2, pp. 131–175.
- Joyce, Terry (2023). "Typologies of writing systems." In: *The Cambridge handbook of historical orthography*. Ed. by Marco Condorelli and Hanna Rutkowska. Cambridge: Cambridge University Press, pp. 138–160.
- Joyce, Terry and Dimitrios Meletis (2021). "Alternative criteria for writing system typology. Cross-linguistic observations from the German and Japanese writing systems." In: *Zeitschrift für Sprachwissenschaft* 40.3, pp. 257–277.
- Klages, Ludwig (1917). *Handschrift und Charakter. Gemeinverständlicher Abriss der graphologischen Technik*. Leipzig: Johann Ambrosius Barth.
- Logan, H. M. (1973). *The dialect of the life of Saint Katherine. A linguistic study of the phonology and inflections*. Vol. 130. Janua Linguarum. Series Practica. The Hague, Paris: De Gruyter Mouton.
- Massé, Joseph François P. (1863). *Grammatologie française. A series of 50 examination papers*. London: David Nutt.
- McIntosh, Angus (1961). "Graphology and meaning." In: *Archivum Linguisticum* 13, pp. 107–120.
- Meletis, Dimitrios (2015). *Graphetik. Form und Materialität von Schrift*. Glückstadt: Verlag Werner Hülsbusch.
- (2018). "What is natural in writing? Prolegomena to a Natural Grapholinguistics." In: *Written Language & Literacy* 21.1, pp. 52–88.
- (2020). *The nature of writing. A theory of grapholinguistics*. Vol. 3. Grapholinguistics and Its Applications. Brest: Fluxus Editions.
- (2021a). "On being a grapholinguist." In: *Grapholinguistics in the 21st Century—2020, Proceedings*. Ed. by Yannis Haralambous. Vol. 4. Grapholinguistics and Its Applications. Brest: Fluxus Editions, pp. 125–141.
- (2021b). "Structural, psycholinguistic, and sociolinguistic typologies of writing." Paper presented at the workshop 'Writing. System, use, ideology' as part of the 46th Austrian Linguistics Conference. Vienna.
- (in press). "There had already been a structural graphemics. Revisiting and contextualizing a grapholinguistic dispute." In: *LACUS Forum* 47.
- Meletis, Dimitrios and Christa Dürscheid (2022). *Writing systems and their use. An overview of grapholinguistics*. Vol. 369. Trends in Linguistics. Studies and Monographs. Berlin, Boston: De Gruyter.

- Mieroop, Marc van de (2021). "On Babylonian grammatology." In: *Signs—Sounds—Semantics. Nature and transformation of writing systems in the Ancient Near East*. Ed. by Costa Gabriel, Karenleigh A. Overmann, and Annick Payne. Vol. 13. Wiener Offene Orientalistik. Münster: Ugarit, pp. 161–169.
- Mihm, Arend (2016). "Zur Theorie der vormodernen Orthographien. Straßburger Schreibsysteme als Erkenntnisgrundlage." In: *Sprachwissenschaft* 41.3–4, pp. 271–309.
- Neef, Martin (2005). *Die Graphematik des Deutschen*. Vol. 500. Linguistische Arbeiten. Berlin, Boston: De Gruyter.
- (2015). "Writing systems as modular objects. Proposals for theory design in grapholinguistics." In: *Open Linguistics* 1, pp. 708–721.
- (2021). "Schriftlinguistik." In: *Schriftlinguistik/Grapholinguistics*. Ed. by Martin Neef, Said Sahel, and Rüdiger Weingarten. Vol. 5. Wörterbücher zur Sprach- und Kommunikationswissenschaft. [https://www.degruyter.com/database/WSK/entry/wsk\\_id\\_wsk\\_artikel\\_artikel\\_13140/html](https://www.degruyter.com/database/WSK/entry/wsk_id_wsk_artikel_artikel_13140/html). Boston, Berlin: De Gruyter.
- Nerius, Dieter (1986). "Zur Begriffsbestimmung im Bereich der geschriebenen Sprache." In: *Wissenschaftliche Zeitschrift der Wilhelm-Pieck-Universität Rostock* 35.8, pp. 36–40.
- (1988). "Einleitung." In: *Probleme der geschriebenen Sprache. Beiträge zur Schriftlinguistik auf dem XIV. internationalen Linguistenkongreß 1987 in Berlin*. Ed. by Dieter Nerius and Gerhard Augst. Vol. A 173. Linguistische Studien. Berlin: Akademie der Wissenschaften der DDR, pp. 1–3.
- (2012). "Zur Geschichte der Schriftlinguistik in der Germanistik der DDR." In: *Positionen der Germanistik in der DDR. Personen – Forschungsfelder – Organisationsformen*. Ed. by Jan Cölln and Franz-Josef Holznagel. Berlin, Boston: De Gruyter, pp. 387–397.
- Nerius, Dieter and Gerhard Augst, eds. (1988). *Probleme der geschriebenen Sprache. Beiträge zur Schriftlinguistik auf dem XIV. internationalen Linguistenkongreß 1987 in Berlin*. Vol. A 173. Linguistische Studien. Berlin: Akademie der Wissenschaften der DDR.
- Nunberg, Geoffrey (1990). *The linguistics of punctuation*. Stanford: CSLI Publications.
- Pettersson, John S. (1996). *Grammatological studies. Writing and its relation to speech*. Vol. 29. Reports from Uppsala University Linguistics. Uppsala: University of Uppsala.
- Piirainen, Ilpo Tapani (1986). "Autonomie der Graphematik in historischer Hinsicht / The autonomy of graphemics from historical point of view." In: *New trends in graphemics and orthography*. Ed. by Gerhard Augst. Boston, Berlin: De Gruyter, pp. 97–104.
- Platt, Penny (1977). "Grapho-Linguistics. Children's drawings in relation to reading and writing skills." In: *The Reading Teacher* 31.3, pp. 262–268.

- Powell, Alexander et al. (2007). "Disciplinary baptisms. A comparison of the naming stories of genetics, molecular biology, genomics, and systems biology." In: *History and Philosophy of the Life Sciences* 29.1, pp. 5–32.
- Pulgram, Ernst (1951). "Phoneme and grapheme. A parallel." In: *WORD* 7, pp. 1–20.
- Sariti, Anthony W. (1967). "Chinese grapholinguistics." MA thesis. Georgetown University.
- Sebba, Mark (2009). "Sociolinguistic approaches to writing systems research." In: *Writing Systems Research* 1.1, pp. 35–49.
- Spitzmüller, Jürgen (2013). *Graphische Variation als soziale Praxis. Eine soziolinguistische Theorie skripturaler ‚Sichtbarkeit‘*. Vol. 56. Linguistik – Impulse & Tendenzen. Berlin, Boston: De Gruyter.
- Stark, Elisabeth (2022). "Warum es nur eine Linguistik gibt. Keine Interdisziplinarität ohne starke Disziplinen." In: *Brückenschläge – Linguistik an den Schnittstellen*. Ed. by Sarah Brommer, Kersten Sven Roth, and Jürgen Spitzmüller. Vol. 583. Tübinger Beiträge zur Linguistik. Tübingen: Narr Francke Attempto, pp. 19–38.
- Tan, Samuel K. (1982). *Selected essays on the Filipino muslims*. Marawi City: University Research Center, Mindanao State University.
- Venezky, Richard L. (1970). *The structure of English orthography*. Vol. 82. Janua Linguarum. Series Minor. The Hague: Mouton De Gruyter.
- Wales, Katie (2014). *A dictionary of stylistics*. 3rd ed. Oxon, New York: Routledge.
- Watt, W. C., ed. (1994a). *Writing systems and cognition*. Vol. 6. Neuropsychology and Cognition. Dordrecht: Springer.
- (1994b). "Foreword." In: *Writing systems and cognition*. Ed. by W. C. Watt. Vol. 6. Neuropsychology and Cognition. Dordrecht: Springer, pp. vii–xiv.
- Zaefferer, Dietmar (2006). "Realizing Humboldt's dream. Cross-linguistic grammatography as data-base creation." In: *Catching language. The standing challenge of grammar writing*. Ed. by Felix K. Ameka, Alan Dench, and Nicholas Evans. Vol. 167. Trends in Linguistics. Studies and Monographs. Berlin, New York: De Gruyter, pp. 113–135.
- Žagar, Mateo (2007). *Grafolingvistika srednjovjekovnih tekstova*. Zagreb: Matica Hrvatska.
- (2020). "Orthographic solutions at the onset of early modern Croatian." In: *Advances in historical orthography, c. 1500–1800*. Ed. by Marco Condorelli. Cambridge: Cambridge University Press, pp. 176–190.
- Zhong, Yurou (2019). *Script revolution and literary modernity, 1916–1958*. New York: Columbia University Press.

# Amodal Morphology. Applications to Brahmic Scripts and Canadian Aboriginal Syllabics


Amalia E. Gnanadesikan

*Abstract.* The discovery of grammar in sign language in the late twentieth century led to the realization that grammar is amodal. Increasingly, writing is being considered a third (admittedly derivative) modality of language, with written signs possessing grammar. Most such work has thus far focused on phonological structures such as graphic syllables and feet, but this paper argues that written signs also have morphology. Morphological analyses of Chinese characters (Hànzi) and of Maya glyph blocks are cited, and new analyses of Brahmic scripts and Canadian Aboriginal Syllabics are presented. Just as the modality of sign languages affects the expression of their grammar, the written modality of scripts affects their grammatical expression. Specifically, they are designed to be processed spatially and as such have some morphological characteristics in common with sign languages. The consonant and vowel schemas of the Canadian Aboriginal Syllabics family of scripts are interpreted here as analogs of the so-called ion-morphs of American Sign Language.

## 1. Introduction

The first two decades of the twenty-first century found an increasing number of scholars (e.g., Baroni (2015), Meletis (2020), Myers (2019), and Primus (2004)) arguing that writing is a true—if derivative—modality of language and that scripts have grammars that can insightfully be analyzed with linguistic tools.<sup>1</sup> This is despite obvious differences between writing and primary language such as the universality and automatic acquisition of primary language and the relatively recent invention of writing (for summaries of the arguments against writing being language see, e.g., Daniels (2018, pp. 183–187) and Gnanadesikan (2021b, pp. 106–108)). In aiming to resolve the tension between the differences between primary

---

Amalia E. Gnanadesikan  0000-0003-4371-9288  
478 Blackshire Road, Severna Park, MD 21146, USA  
E-mail: amalia.gnanadesikan@gmail.com

1. Thanks to Daniel Harbour, Dimitrios Meletis and the anonymous abstract reviewers for discussions relating to this topic, and to Yannis Haralambous for creating the venues to share and publish this and other work in grapholinguistics.

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 45–66. <https://doi.org/10.36824/2022-graf-gnan>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

language and writing systems on the one hand and their similarities on the other, I have elsewhere argued that “language is indeed cognitively special but that this specialness lies not so much in being unique but in being overpowering” (Gnanadesikan, 2021b, p. 118), with the result that over time writing becomes language, both phylogenetically (historically) and ontogenetically (in the life of a literate person).

The argument for writing as a modality of language has been made plausible by a foundational discovery of the twentieth century, namely that the human capacity for language is essentially amodal. That is, (primary) language can be either signed or spoken, with signed languages having the same complexity of grammar as that found in spoken language. The discovery of complex grammar in sign language was not a foregone conclusion for researchers at the time. In an admission that is startling to read in hindsight, Edward Klima and Ursula Bellugi state in their classic book *The Signs of Language* that they “did not begin by assuming that A[merican] S[ign] L[anguage] had a grammar...” (Klima and Bellugi, 1979, p. 3). However, “American Sign Language turned out to be in fact a complexly structured language with a highly articulated grammar, a language that exhibits many of the fundamental properties linguists have posited for all languages. But the special forms in which such properties are manifested turn out to be primarily a function of the visual-gestural mode” (ibid., p. 4).

Once the amodal nature of language is accepted and its implications absorbed, the discovery of grammatical structures in scripts should perhaps not be surprising. Indeed, as James Myers has stated, “[A]n amodal capacity for grammar would not only explain sign languages but would also predict that grammar should appear beyond both speech and sign...” (Myers, 2019, p. 22). However, just as signed languages were found to have special properties because of the visual-gestural mode in which they operate, written languages can be expected to be influenced by their modality in their choice and expression of grammatical structures.

Searching for written correlates of grammatical structures found in spoken languages has been a productive line of research in recent years. Thus a number of studies have found correlates of phonological structures in writing systems, such as graphic features (Primus, 2004), graphematic syllables (Fuhrhop, Buchmann, and Berg, 2011; Myers, 2021), and graphematic feet (Evertz, 2018; Evertz and Primus, 2013). This paper argues that there are also correlates of morphological structures in writing systems. In making this argument I follow Beatrice Primus, who has described the letters of the Roman alphabet as “made of syntagmatically concatenated smaller units. Loosely speaking, they resemble morphemes rather than phonemes” (Primus, 2004, p. 240). A second point of resemblance between written signs (or graphemes) and morphemes is that they are both signs, with a signifier and a signified, unlike phonemes, which are meaningless except in combination

(Meletis, 2020, p. 202). As a result, finding morphological structure in written signs should perhaps not be too surprising. In this paper I argue that not only are there correlates of spoken morphological structures to be found in the structure of written signs, but there are also correlates of specifically sign language structures. This claim also stands to reason, given both the amodality of language and the shared visual nature of the signed and written modalities.

Before continuing, it is best to consider briefly what is meant by morphology in spoken language. Morphology as a study is “the branch of linguistics that deals with words, their internal structure, and how they are formed” (Aronoff and Fudeman, 2011, p. 1). As part of the mental grammar, it is “the mental system involved in *word* formation” (ibid., 1, bold in original). Morphemes can be stems or affixes. Affixes include prefixes, suffixes, infixes, and circumfixes. Morphology can be inflectional (which produces a word whose basic meaning is not changed from the core meaning of the stem) or derivational (which does change the core meaning of the stem and may change the lexical category of the resulting word). Derivational morphology may be accomplished via compounding, affixation, or zero derivation. Other minor ways to derive new lexemes include blending/portmanteaus, acronyms, clipping, and backformation (ibid.). In considering the existence of morphological analogs in writing, the important abstraction to make is to replace the concept of *word* in the foregoing paragraph with *sign*. Words are signs, with a form and a meaning, and often with internal structure (morphology). Written signs are also signs, with a form and an interpretation, or meaning—and often with internal structure. Once the abstraction of *word* to *sign* made, the morphological analysis of written signs can begin.

The rest of this paper is structured as follows. Section 2 reviews Myers’ (2019) analysis of morphemic structure in Chinese characters, while Section 3 reviews the conventional analysis of the structure of Maya glyph blocks, which can also be seen as being essentially morphemic. Section 4 applies a morphemic analysis to the scripts of the Brahmic family, arguing for stem-and-affix structures and compounds in the complex aksharas of those scripts. Section 5 presents the concept of ion-morphs in American Sign Language formulated by Fernald and Napoli (2000) and applies this concept to the analysis of scripts in the Canadian Aboriginal Syllabics family, notably Carrier. Section 6 concludes.

## 2. Previously Identified Morphology in Script: Chinese characters

The characters of Chinese script (Hànzi) are famous for encoding syllable-sized morphemes in Chinese (Myers, 2019, p. 2), leading to typological classifications of the script as morphographic (Daniels, 2018,

p. 85) or morphosyllabic (DeFrancis, 1989, p. 115).<sup>2</sup> Most characters in the script are complex, usually composed with a so-called radical or semantic component, which provides a clue to the semantic field of the morpheme, and a phonetic component, which gives a clue to the pronunciation of the syllable. Myers (2019) analyzes these complex Hànzì characters as having morphemic structure, with the semantic radicals being akin to affixes.

For example, the character < 妈 > *mā* ‘mother’ (using here simplified characters as used in mainland China) has two components although the character as a whole represents a monomorphemic word. One component, the so-called radical, is < 女 > *nǚ* ‘female’, which gives a clue to the semantic class of the represented morpheme. The other component is < 马 > *mǎ* ‘horse’, which gives a clue to the pronunciation of the syllable. Such composite characters are usually referred to as semantic-phonetic compounds, but as Myers analyzes them, they are structurally composed of a stem and affix. The semantic radical plays the role of an affix (or affix-like morpheme correlate, despite the apparent implications of root-hood inherent in the traditional term *radical*), as it belongs to a closed class of signs, is semantically bleached, and often occurs in reduced, bound form (ibid., § 2.3.1). A composite character can itself form the phonetic component of another character. In other words, an affixed character may take further affixes, just as a stem-affix structure in a spoken language may take additional affixes. Myers identifies the semantic radicals/affixes as inflectional on the grounds that they are the only character components that display any (although limited) agreement: in the rare cases where a morpheme is disyllabic and represented with two characters, the radicals of the two characters will match (p. 64).

Myers also identifies compounds among complex characters, such as < 晃 > *gǎo* ‘bright’, which is composed of < 日 > *rì* ‘sun’ and < 木 > *mù* ‘tree’. In such a case, neither one of the character components is reduced with respect to the other, and they belong to a more open class and display less semantic bleaching than semantic radicals. Thus Myers identifies such composites as compounds (ibid., § 2.3.2). Another morphological structure that Myers identifies in composite characters is reduplication, for example in < 多 > *duō* ‘many’, which is composed of two instances of < 夕 > *xī* ‘evening’ (ibid., § 2.3.3).

The morphological analysis of Hànzì characters is entirely independent of both the morphological nature of the Chinese language (which in fact employs very little inflectional morphology) and of the morphographic nature of how the script encodes the spoken language. Rather, bi- or polymorphemic characters stand for lexical items which have

---

2. As used in Japanese, however, the characters encode morphemes that are not necessarily syllable-sized. In the Japanese use of kanji, the characters more clearly represent morphemes rather than syllables (Joyce, 2011).



no internal morphological structure. Thus the morphological structure here is entirely a characteristic of the *script* and implies nothing about the *language* (in which complex morphological structures, usually compounds, are represented with sequences of characters). For clarity I refer to the graphic analog of morphology in the script as G-morphology henceforth and the lexical morphology in the language as L-morphology.

As the use of inflectional morphology in Hànzì demonstrates, the presence of a grammatical feature in G-morphology does not depend on its existence in the L-morphology of the corresponding spoken language. In other words, the human instinct for grammar directly influences the form of writing systems independently of its influence on primary (spoken or signed) language.

### 3. Previously Identified Morphology in Script: Maya glyph blocks

The classic Maya script is an extinct morphosyllabic script of Mesoamerica whose signs are conventionally referred to as *glyphs*. The individual signs may stand for lexical morphemes (of shape CVC or CVCVC) or for syllables (of shape CV), which may function as phonetic complements or as the sole spelling of either lexical morphemes or grammatical affixes. A single sign or two or more signs combined together form a *glyph block*, whose major constituent is known as the *main sign* (Law and Stuart, 2017, pp. 130–133). Smaller, narrower signs arranged around the edges of a main sign are conventionally known as affixes (Montgomery, 2002, pp. 43–45), though perhaps with no theoretical grapholinguistic intent. Glyph blocks are often, but not always, arranged in double columns (from left to right) in texts.

The smaller signs known as affixes in the Maya script may occur to the left, right, top, or bottom of the main sign and are known as prefixes, postfixes, superfixes, and subfixes accordingly. The order of reading is usually (but not invariably) prefix, superfix, main sign, postfix, subfix. Glyph blocks may contain only syllabic signs, only morphographic signs, or (as is often the case) some combination of the two. Figures 1 through 3 show examples of the composition of glyph blocks.

If we consider the main signs to be graphic stems and accept the affixes as true graphic affixes, then a morphological analysis of Maya glyph blocks is already done. There is some relationship between the L-morphology of Classic Mayan and the G-morphology of glyph blocks in that “a single glyph block rarely contains incomplete portions of two different morphemes” (Law and Stuart, 2017, p. 130). However, a single word can span more than one glyph block, and a single glyph block may

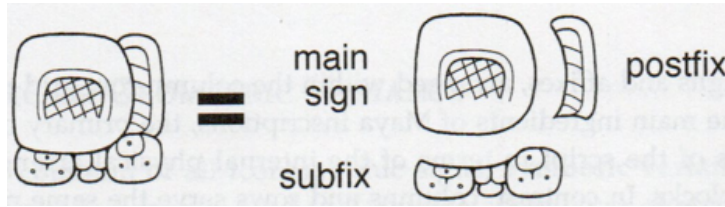


FIGURE 1. An example of a Maya glyph block composed of three syllabograms (from Montgomery (2002, p. 43), used with permission). The main sign reads /pa/, the postfix reads /ka/, and the subfix reads /la/, for a complete reading of *pakal* 'shield'.

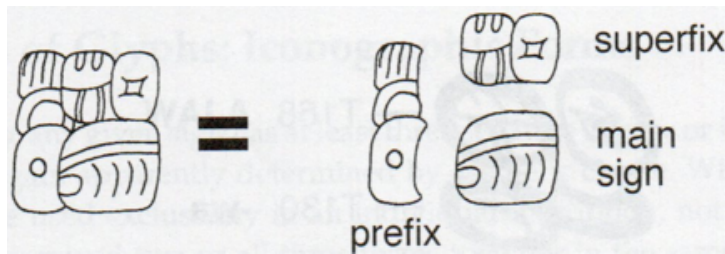


FIGURE 2. An example of a Maya glyph block composed of two syllabograms and a morphogram (from Montgomery (2002, p. 43), used with permission). The prefix reads /ti/, the superfix reads *AJAW* (a morphogram), and the main sign reads /le/, for a complete reading of *ti ajawle(l)* 'in office'.

contain a whole phrase (as in Figure 2). The main sign is often a morphogram (as in the left two examples in Figure 3) but it may also be a syllabogram (as in Figure 1 and 2 and the rightmost example in Figure 3). The G-affixes are often syllabograms but can also be morphograms, especially numerals. Significantly, main signs (G-stems) do not necessarily represent L-stems and the G-affixes do not necessarily represent L-affixes. In other words, the G-morphology is not a mere reflection of the L-morphology. Instead, it is an analog of morphology, independently produced by the human language faculty.

What is not so clearly analogous to L-morphology in the structure of Maya glyph blocks is the presence of superfixes and subfixes in addition to the more familiar prefixes and postfixes (which could just as easily have been called suffixes). However, the addition of more affix types is merely the expression of affixation in two-dimensional space. Unlike a spoken signal, which unfolds in one-dimensional, unidirectional time, a written message is two dimensional. Rather than considering the different structures that arise from differences in modality between speech and glyphs as evidence against the linguistic nature of scripts, we should



FIGURE 3. Three ways to write *usiiij* ‘vulture’ in Maya glyphs, with a simple glyph block on the left, a complex glyph block with a morphographic main sign and three syllabic phonetic complements as G-affixes in the middle, and a complex glyph block with purely syllabic spelling on the right (from Law and Stuart (2017, p. 131), used with permission).

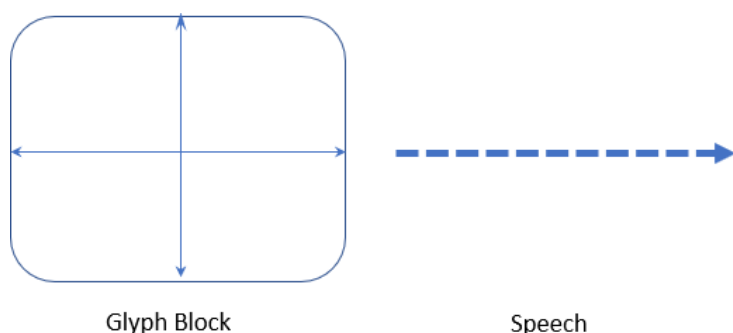


FIGURE 4. A schematic representation of the two-dimensional arrangement of Maya glyph blocks as compared to the one-dimensional, unidirectional arrangement of speech

in fact expect a difference in modality to affect the expression of grammar, as it does in the difference between signed and spoken language.

Glyph blocks may also display morphological structures that are less common but still well known from L-morphology: fusion, analogous to blends in L-morphology (like *smog* or *brunch*), and infixation. A glyph block composed with infixation is shown in Figure 5.

The two writing systems discussed in this section and the previous one are heavily morphographic, with Hànzì characters usually representing monosyllabic morphemes (when used in Chinese) and individual Maya glyphs representing either a syllable or a morpheme. Although the G-morphological structures of the written signs presented in this section do not reflect the L-morphological structure of the words of the relevant languages, it is perhaps no great stretch for readers of these scripts to think of the structure of the complex signs in these scripts in



FIGURE 5. Infixation in Maya glyph blocks. The main sign is a morphogram reading *CHUM* and the infix is a syllabographic phonetic complement reading /mu/ for a full reading of *chum* ‘zero’ (from Montgomery (2002, p. 42), used with permission).

terms of morphology—as witnessed by the traditional terminology of affixes in Maya glyph blocks. The next two sections, however, expand the concept of G-morphology to more fully phonographic scripts—the Brahmic scripts of South and Southeast Asia and the Canadian Aboriginal Syllabics script family of North America—where morphology is perhaps less expected but arguably just as present.

#### 4. Morphology in Script: Brahmic Scripts

The Brahmic scripts are a family of scripts used in South and South-east Asia which includes Devanagari (used for Hindi, Nepali, and Marathi), Bengali, Gujarati, Oriya, Gurmukhi (used for Punjabi), Tibetan, Tamil, Kannada, Telugu, Malayalam, Sinhalese, Burmese, Thai, Lao, and Khmer, as well as numerous unofficial and/or historical scripts in the area. Although there is a certain amount of variation in the design and use of these scripts, they share descent from the ancient Brāhmī script and the design feature of the akshara as an organizing unit. Simple, primary signs (aksharas) either represent consonants, often accompanied by a default or “inherent” vowel, or initial vowels. Vowels other than the default vowel (when not initial) are written as satellites above, below, to the right, to the left, or surrounding the consonantal sign (Gnanadesikan, 2021a).

Table 1 illustrates Devanagari as used for modern Hindi (Bright, 1996b; Snell and Weightman, 1989). The first four rows list the simple aksharas, with the independent vowel letters (V) in the first row and the consonants (Ca, with inherent vowel) in the second through fourth rows. The fifth row illustrates the addition of satellite vowel signs to consonants to form complex CV aksharas, and the sixth row illustrates the combination of consonants to form complex CCa aksharas. The final row illustrates complex aksharas that both combine consonants and add a satellite vowel sign (CCV).

Considering the structure of Devanagari and related scripts through the lens of morphology, the consonants (and independent vowels) can

TABLE 1. Devanagari as used for modern Hindi, accompanied by the conventional transliteration. The first four rows list the simple aksharas of the script while the last three rows give a sample of complex aksharas that include satellite vowels, conjunct consonants or both.

|      |       |      |      |      |      |      |      |      |      |      |
|------|-------|------|------|------|------|------|------|------|------|------|
| अ    | आ     | इ    | ई    | उ    | ऊ    | ऋ    | ए    | ऐ    | ओ    | औ    |
| a    | ā     | i    | ī    | u    | ū    | ṛ    | e    | ai   | o    | au   |
| क    | ख     | ग    | घ    | ङ    | च    | छ    | ज    | झ    | ञ    |      |
| ka   | kha   | ga   | gha  | ṅa   | ca   | cha  | ja   | jha  | ña   |      |
| ट    | ठ     | ड    | ढ    | ण    | त    | थ    | द    | ध    | न    |      |
| ṭa   | ṭha   | ḍa   | ḍha  | ṇa   | ta   | tha  | da   | dha  | na   |      |
| प    | फ     | ब    | भ    | म    | य    | र    | ल    | व    | श    | ष    |
| pa   | pha   | ba   | bha  | ma   | ya   | ra   | la   | va   | śa   | ṣa   |
| स    | ह     |      |      |      |      |      |      |      |      |      |
| sa   | ha    |      |      |      |      |      |      |      |      |      |
| क    | का    | कि   | की   | कु   | कू   | कृ   | के   | कै   | को   | कौ   |
| ka   | kā    | ki   | kī   | ku   | kū   | kṛ   | ke   | kai  | ko   | kau  |
| व्य  | क्र   | क्ल  | ल्क  | ट्ट  | प्र  | श्ल  | स्क  | स्व  | स्म  | न्य  |
| bya  | kra   | kla  | lka  | ṭṭa  | pna  | śla  | ska  | sva  | sma  | nya  |
| ष्ट  | त्त्व |      |      |      |      |      |      |      |      |      |
| ṣṭa  | ttva  |      |      |      |      |      |      |      |      |      |
| व्या | क्री  | क्ले | ल्कु | ट्टू | प्नि | श्लो | स्कृ | स्वै | स्मा | न्यौ |
| byā  | krī   | kle  | lku  | ṭṭū  | pni  | ślo  | skṛ  | svai | smā  | nyau |
| ष्टि | च्छे  |      |      |      |      |      |      |      |      |      |
| ṣṭi  | cche  |      |      |      |      |      |      |      |      |      |

be seen as stems and the satellite vowels as affixes. Because of the two-dimensional nature of the written modality, the vowel affixes can appear anywhere around the consonant stems, including above or below them (as in <के> /ke/ and <कु> /ku/), not just before or after the stems as in spoken prefixes and suffixes. In other words, the structure of Devanagari aksharas resembles the structure of Maya glyph blocks, with a main sign and optional prefixes, superfixes, postfixes, and subfixes.

Can we assign the affixes to the category of inflectional or derivational? Arguably, yes. Because the vowel affixes do not alter the fundamental identity of the consonantal main sign, or stem, we can consider the satellite vowel affixes to be inflectional affixes. Considered in this way, the inherent vowel is analogous to a default inflection which may occur without the addition of an affix in L-morphology to express categories such as singular number or nominative case. Thus, rather than being a typological oddity, the inherent vowel is in good morphological company.

Derivational morphology may be seen in another aspect of the script. In order to write phonemes that were not part of the Sanskrit inventory but which have been added as the result of historical change or lexical borrowing, Hindi adds a diacritic dot. Since this dot alters the identity of the consonant, it can be considered a derivational affix.

TABLE 2. Derivational morphology in Devanagari. Adding the diacritic dot to the aksharas on the left produces those on the right, with changes to the value of the consonants.

|   |     |    |                          |
|---|-----|----|--------------------------|
| क | ka  | क़ | qa                       |
| ख | kha | ख़ | kha (/xa/)               |
| ग | ga  | ग़ | ga (/ɣa/)                |
| ज | ja  | ज़ | za                       |
| ड | ḍa  | ड़ | ṛa (/ɽa/)                |
| ढ | ḍha | ढ़ | ṛha (/ɽ <sup>h</sup> a/) |
| फ | pha | फ़ | fa                       |

Devanagari also provides examples of compounding. These are the cases in the last two rows of Table 1, in which two or more consonants are represented by a complex akshara. When, for example, <ब> /ba/ and <य> /ya/ combine, the result is <ब्य> /bya/. The left-hand member of the consonant conjunct is usually somewhat reduced (a notable exception being conjuncts ending with <र> /ra/, which reduce the right-hand member), yielding what we can describe morphologically as a right-headed compound. As expected under this analysis, the left-hand (non-head) member does not receive default inflection (the inherent vowel) but the right-hand member (the head) does if it does not take an inflecting affix (i.e., vowel sign). Thus there is a single default vowel for the conjunct, rather than two.

Other Brahmic scripts provide examples of other morphological structures and processes. In the Kannada script, used for the Kannada language of South India, the conjunct consonants form left-headed rather than right-headed compounds. When two consonants are joined to form a complex akshara in Kannada, the first one remains full size and is the attachment point for satellite (affixed) vowels, while the second one is reduced (Bright, 1996a). Many conjunct consonants represent geminate (double) consonants, but when the two consonants are distinct, it is clear that it is the second one that is reduced and the first one to which the vowel is attached.<sup>3</sup> Examples are shown in Table 3, where the first row displays some simple consonantal aksharas (with default vowel), the second line shows those same consonants with an affixed vowel (the /e/ vowel is a small curl added at the top of the consonant), and the third line shows conjunct consonants, with and without an affixed vowel.

3. As is commonly the case in Brahmic scripts, consonant clusters of which the first element is /r/ are written differently than other clusters, so conjuncts representing rC clusters in Kannada are an exception to the left-headed generalization given above (Bright, 1996a).

TABLE 3. Examples of simple (top row) and complex aksharas in Kannada. The middle line shows CV aksharas, while the third line shows aksharas with conjunct consonants.

|     |     |      |           |     |           |      |           |
|-----|-----|------|-----------|-----|-----------|------|-----------|
| ತ   | ta  | ಯ    | ya (/ja/) | ಜ   | ja (/ɟa/) | ಞ    | ña (/ɲa/) |
| ತೆ  | te  | ಯೆ   | ye        | ಜಾ  | jā        | ಞಾ   | ñā        |
| ತ್ಯ | tya | ತ್ಯೆ | tye       | ಜ್ಞ | jña       | ಜ್ಞಾ | jñā       |

TABLE 4. The sign <ಕ> /ka/, shown alone (top row, far left) and with each satellite vowel added

|    |    |     |    |    |     |
|----|----|-----|----|----|-----|
| ಕ  | ಕಾ | ಕಿ  | ಕೀ | ಕು | ಕು  |
| ka | kā | ki  | kī | ku | kū  |
| ಕೆ | ಕೇ | ಕಾ  | ಕೊ | ಕೋ | ಕೋ  |
| ke | kē | kai | ko | kō | kau |

In the Tamil script, used for the Tamil language of South India and Sri Lanka, we see circumfixes, the operation of stem-conditioned allomorphy, and two distinct levels of affixation. In Tamil script there are no conjunct consonants, unlike in Devanagari (or Kannada). The vowels are, as in Devanagari, added to the left, to the right, above, or below the consonant, with the additional detail that some satellite vowels have two parts, one of which appears to the left and one of which appears to the right of the main consonant sign. The sign <ಕ> /ka/, with the addition of the various vowels signs, is shown in Table 4 (Steever, 1996).

As shown at the bottom right in Table 4, the signs for the vowels /o/, /ō/ and /au/ are circumfixes, with one part of the vowel written before and one part written after the consonant main sign. Circumfixed vowels are common in Brahmic scripts, occurring in Oriya, Bangla, Malayalam, Sinhala, Burmese, Khmer, Lao, and Thai (Gnanadesikan, 2021a). Some other Brahmic scripts (such as Kannada) that do not place any vowels to the left of a consonant still have two-part vowels, with one part written above and one part written to the right of a consonant. Circumfixes are rare in L-morphology, and it has been suggested that they are best analyzed as a simultaneous application of a prefix and a suffix. Whatever their ideal analysis, however, they do occasionally occur (Aronoff and Fudeman, 2011, p. 3). And whatever their ideal analysis (two morphemes or one), an analogous analysis can be applied to the Tamil circumfixed vowels. The fact that circumfixed vowels are quite common in Tamil and many other Brahmic scripts can be seen as a case of modality influencing the expression of grammar. Writing is not as unidirectional as speech, in that a sign or word that has been written or read is still present when the next word or sign is being written or read. Lookback is thus easily accomplished in the written modality, unlike in speech.

Easy lookback would understandably make disjoint signs less dispreferred than they are in speech. Thus G-morphology presents us with structures familiar from spoken, lexical morphology but with added options for affix placement (including above and below) and greater comfort with disjoint signs

Another feature of Tamil aksharas is allography of vowel signs triggered by the shape of the consonant, which in a morphological analysis is stem-conditioned allomorphy. The satellite sign for the vowel /u/ has three different allographs, depending on the consonant sign (G-stem) to which it attaches. As shown in Table 5, the sign representing the vowel /u/ can be a curve that starts on the bottom right of the main sign, passes under the main sign and emerges on the left. Or it can start on the bottom right of the main sign, pass under the main sign and reverse course to emerge back on the right. A third option is for it to be simply a short vertical line under the right-hand side of the main sign. The Cu aksharas in Table 5 are accompanied by their unaffixed Ca stems in parentheses to allow comparison of the affixed and unaffixed forms. While it is difficult to state fully and exactly the properties of the stems that condition the allomorphy, some generalizations are clear. The downward curving tail on <த> /ta/, <ந> /na/, and <ற> /ra/ triggers the reversing allomorph (second row). Stems whose right edge is a vertical line without an overhang (i.e., <ங> /ṅa/, <ப> /pa/, <ய> /ya/, and <வ> /va/, but not <ன> /na/ due to the overhang) all share the subfixed vertical line (third row).

TABLE 5. Allography in the satellite sign for the vowel /u/ in Tamil. The basic Ca aksharas are shown in parentheses after each Cu akshara for comparison.

|                     |                      |              |                     |              |                     |              |
|---------------------|----------------------|--------------|---------------------|--------------|---------------------|--------------|
| கு (க)<br>ku        | டு (ட)<br>tu (/tu/)  | மு (ம)<br>mu | ரு (ர)<br>ru        | லு (ழ)<br>lu | லு (ள)<br>lu (/lu/) |              |
| நு (ந)<br>ṅu (/ṅu/) | னு (ண)<br>ṇu (/ṇu/)  | து (த)<br>tu | நு (ந)<br>nu (/ṇu/) | லு (ல)<br>lu | ரு (ற)<br>ru        | னு (ன)<br>nu |
| பு (ப)<br>pu (/pu/) | சு (ச)<br>cu (/tʃu/) | பு (ப)<br>pu | யு (ய)<br>yu (/ju/) | வு (வ)<br>vu |                     |              |

The allography in Tamil script shown in Table 5 is analogous to allomorphy in affixation such as the English plural allomorphy exemplified by *dogs*, *cats*, and *horses*, in which the English plural takes the form [s], [z] or [əz] depending on the voicing and/or sibilance of the final segment of the stem. Like the English plural, the Tamil script /u/ affix is sensitive to the shape of the stem.

The G-affix representing /ū/ has similar allomorphy. Other vowel affixes have historically had allomorphy as well, but it has largely been eliminated. It is significant that the affixes that continue to display al-



lography/allomorphy are ones that are tightly connected to their stems. Tamil vowel affixes can be divided into two groups: those that are graphically connected to their stems and those that have a space. Looking back at Table 4, it can be seen that the affixes for /u/, /ū/, /i/ and /ī/ are tightly connected to the main sign while the rest do not touch the main sign. While it may be tempting—because of the gap—to consider these latter vowel signs as something other than proper affixes, they are nevertheless dependent, bound signs that do not occur without an accompanying consonantal main sign (with the exception of the second half of the <௮௫> /au/ sign, which happens to be homologous with the simple akshara ஂ /a/). I would argue that the difference between the two types of vowel signs can be accounted for by considering the Tamil script as containing two levels of affixes. This division of the morphology into two levels is a phenomenon familiar from English morphology, in which Level 1 (or primary) affixes such as the prefix *in-* occur closer to the stem and involve more allomorphy than Level 2 (or secondary) affixes such as the prefix *un-* (Aronoff and Fudeman, 2011, p. 86). In Tamil script, the /u/, /ū/, /i/ and /ī/ signs are the tightly bound Level 1 affixes, and the others are the less tightly bound Level 2 affixes.

In summary, the Brahmic scripts provide many opportunities to see analogs of morphology in the structure of written signs, including stem-affix structures, stem-conditioned affixal allomorphy, both derivational and inflectional affixes, both left- and right-headed compounds, and a division of the affixes into two levels.

## 5. Ion-Morphs in American Sign Language and Carrier Syllabics

The analogs to morphology that have been presented in this paper so far have been analogous to the morphology of spoken language. However, if we remember that morphology, like other parts of grammar, is essentially amodal, we should expect to find correlates not just of spoken morphology but of *signed* morphology. This section shows that such analogs do in fact exist.

In signed languages, affixation is rare and other morphological relations dominate (Johnston, 2006). American Sign Language (ASL), for example, has been found to have only one affix (Liddell and Johnson, 2011, p. 329). However, that does not mean that there are no systematic relations between words. One notable property of the lexicon of ASL is the existence of lexical families whose members are related via commonalities in how the signs are made. For example, a lexical family of signs will share two or three of the four manual articulatory parameters (hand configuration, movement, place of articulation, and orientation) while varying one or two of the parameters. Thus the lexical family that includes FAMILY, CLASS, TEAM, GROUP, ASSOCIATION, and SOCIETY

shares movement, orientation, and place of articulation but varies the hand configuration (Fernald and Napoli, 2000). Some of these signs are shown in Figure 6 (by convention, glosses of signs are shown in small capitals).

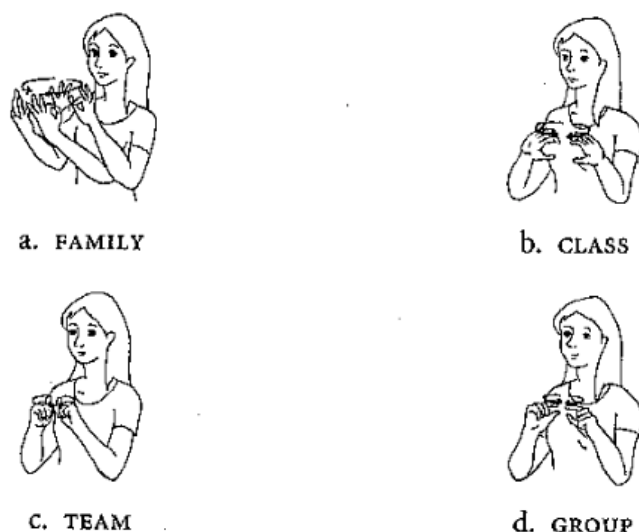


FIGURE 6. Part of the lexical family of signs relating to social groups in ASL (from Fernald and Napoli (2000, p. 20), used with permission)

The members of these lexical families cannot be derived from each other via a process of affixation. In other words, no one member of the lexical family is a basic sign to which material is added to derive the other lexical items in the family. Fernald and Napoli (*ibid.*) analyze these lexical families as being related to each other via the existence of abstract, incomplete lexical units that compose with other such partially specified lexical elements to form complete signs. They term these incomplete units *ion-morphs*. Lexical families share an ion-morph. A sign may belong to multiple families in intersecting dimensions and share ion morphs with more than one set of lexical items. Thus MOTHER and FATHER form a lexical family varying only in place of articulation, but MOTHER is also related to GIRL, sharing with it every parameter except hand configuration, and also to GRANDMOTHER, sharing with it every parameter except movement. Thus MOTHER contains three partially overlapping ion-morphs, sharing one with FATHER, one with GIRL, and one with GRANDMOTHER.

The concept of ion-morphs is somewhat similar to that of templatic morphology (also called root-and-pattern morphology), as seen for ex-

ample in Semitic languages, in which the consonants of a stem are mapped to a template and the vowels, which belong to a different morpheme, are interleaved with the consonants (McCarthy, 1981). So, for example, a root of the shape  $C_1C_2C_3$  might be realized as  $C_1iC_2C_2aC_3$ . However, as Fernald and Napoli (2000) point out, such templates crucially impose an order on consonants and vowels, while the ion-morphs of an ASL sign are articulated simultaneously, without relative order.

The difference between signed and spoken morphology is directly related to the difference in their modality. Fernald and Napoli (*ibid.*) suggest several reasons for the preference for ion-morphs over affixes in ASL. These include the greater bandwidth of visual processing as compared to auditory processing and the relative slowness of the manual articulators as compared to the oral articulators. Taken together, these factors make morphs that are articulated simultaneously preferable to ones articulated in sequence in sign languages. In principle, a language might use neither ion-morphs nor derivational morphology, but as Fernald and Napoli point out, using lexically related signs rather than unrelated signs for lexemes with similar semantic content helps signers who may not know a particular sign to guess its meaning. This is a particular boon to learners of sign languages, who are often (because of being born to hearing parents) older when they begin learning to sign than hearing children are when they begin learning to speak a spoken language.

The concept of ion-morphs provides a key to the analysis of a family of scripts that have not yet found a comfortable home, typologically speaking. *Canadian Aboriginal Syllabics* is a cover term for a family of scripts which have been used for various indigenous languages of Canada, including Cree, Inuktitut, Ojibwe, and Carrier, among others. The first members of the script family were Cree and Ojibwe “syllabics,” invented in 1840 by the missionary James Evans. Adoption, adaptation, and additions in subsequent decades resulted in a family of scripts used to write various languages in the Athabaskan, Algonquian, and Eskimo-Aleut language families Nichols (1996).

Generally speaking, in these scripts the shape of a main sign indicates its consonantal value, and the orientation of the sign (sometimes with an additional diacritic) indicates the value of the following vowel. Smaller signs represent consonants that are not followed by vowels (Nichols, 1996; Poser, 2011). Table 6 shows a selection of the CV signs that are used to write CV sequences in Carrier, an Athabaskan language spoken in central British Columbia, Canada. This particular member of the script family was adapted from earlier variants for the Carrier language by Father Adrien Gabriel Morice in 1885. In Carrier it is known as  $\mathfrak{D}^{\text{v}}\mathfrak{h}\mathfrak{B}$  *dulkw'abke* ‘frog feet’. Initially popular and considered easy to learn, its use declined in the 1920s due to residential English-medium education (Poser, 2003; 2011).

TABLE 6. Some of the CV signs in Carrier “syllabics.” Consonants are represented by the shape of a sign, and vowels are represented by its orientation and/or by the addition of a dot or small line. Voiced, voiceless, and glottalized consonant families are related by modifications of the shape. Adapted from Poser (2003; 2011), using IPA transcriptions rather than the Carrier Linguistic Committee’s Romanization system.

|       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|
| V hu  | < ha  | ^ ho  | > hΛ  | > he  | > hi  |
| U du  | C da  | ∩ do  | ∩ dΛ  | ∩ de  | ∩ di  |
| U tu  | ⌒ ta  | ⌒ to  | ⌒ tΛ  | ⌒ te  | ⌒ ti  |
| ʉ t’u | ⌒ t’a | ⌒ t’o | ⌒ t’Λ | ⌒ t’e | ⌒ t’i |
| ʋ gu  | ε ga  | ∩ go  | ε gΛ  | ε ge  | ε gi  |
| ʋ ku  | ε ka  | ∩ ko  | ε kΛ  | ε ke  | ε ki  |
| ʋ k’u | ε k’a | ∩ k’o | ε k’Λ | ε k’e | ε k’i |

Concentrating first on the first two rows of Table 6, one can see that the shape of a sign represents the consonant value in the CV sequence, and that the sign’s orientation and/or the addition of an internal small line or dot represents the vowel. The Cree language, whose syllabics system inspired the Carrier system, has only four vowel qualities (as well as length), so adapting the system to Carrier required adding extra ways to represent vowels beyond 90-degree rotation.

Carrier also has many more consonants than Cree does, Carrier having 41 native consonants and three further ones that occur only in loanwords, while Cree has ten native consonants and two that occur only in loanwords (Poser, 2011, p. 50). In order to meet the demand for additional consonants, an additional degree of relationship was added to the Carrier script. Comparing the second, third, and fourth rows of Table 6 with the fifth, sixth, and seventh rows, it can be seen that, for a class of plosive consonants that differ only in laryngeal values, the voiced member will be written with a symbol that is open at one end, the voiceless member will be written with a straight closing line, and the glottalized member will be written with a slightly V-shaped closing line.

Other family resemblances exist between the signs used to write other sets of similar consonants. Table 7 shows the signs for CV signs in which the C is a lateral consonant. These signs share a bowl shape with various modifications. An interesting feature revealed by the lateral signs is that the relationship between signs that share a consonant is not purely rotational. Thus the relationship between <ʋ> /t’lu/ and <ε> /t’la/ is not exactly one of a ninety-degree rotation, as the position of the small vertical line that occurs on the top righthand side of /t’lu/ (but also on the top righthand side of /t’la/) demonstrates.

Consonants that are not followed by a vowel are written differently than the main CV signs. The list of consonants that may appear before another consonant or may close a syllable in Carrier is about half as

TABLE 7. The CV signs for syllable initial lateral consonants in Carrier. Adapted from Poser (2003; 2011)

|        |        |        |       |        |        |
|--------|--------|--------|-------|--------|--------|
| U lu   | C la   | Q lo   | ᑭ lᐱ  | ᑭ le   | ᑭ li   |
| U ɬu   | C ɬa   | Q ɬo   | ᑭ ɬᐱ  | ᑭ ɬe   | ᑭ ɬi   |
| U tlu  | C tla  | Q tlo  | ᑭ tᐱ  | ᑭ tle  | ᑭ tli  |
| U dlu  | C dla  | Q dlo  | ᑭ dᐱ  | ᑭ dle  | ᑭ dli  |
| ᑭ t'lu | ᑭ t'la | ᑭ t'lo | ᑭ t'ᐱ | ᑭ t'le | ᑭ t'li |

TABLE 8. Some of the smaller signs that are used for isolated consonants (consonants not followed by vowels). Adapted from Poser (2003; 2011)

|     |     |     |     |     |
|-----|-----|-----|-----|-----|
| ᑭ m | ᑭ n | ᑭ ŋ | ᑭ x | ᑭ ɣ |
|-----|-----|-----|-----|-----|

long as that of those that may occur in CV sequences. These consonants are written with smaller signs which stand only for a single, isolated C. These isolated C signs are unrelated to those that are used to write the same consonant before a vowel. However, even among the isolated Cs, systematic relationship can be observed in the signs for phonologically related consonants, such as when nasals are related by rotation or velar fricatives are related by a change in angle. Some of the isolated consonants are shown in Table 8.

Systematic relationships between signs that represent phonologically related phonemes are a significant feature of the Han'gul script, used for Korean. The representation of phonological features in Han'gul is pervasive enough to have led Sampson (2015) to consider it a “featural system” (representing phonological features of segments) rather than an alphabet (representing phonological segments). While many other grapholinguists disagree with Sampson’s typological placement of Han'gul (Sproat, 2000, pp. 136–138), the script remains justly famous for its inclusion of phonological features in its representation of language. Like Han'gul, the Carrier script also represents relationships between phonological segments (in other words, phonological features) with systematic relationships between signs (such as the line added to <U> /du/ to derive <ᑭ> /tu/). However, not all of the relationships between signs can be reduced to the addition of something (a stroke or dot, say) to a simpler shape. Specifically, the vowels are not represented by *adding* anything. Instead, the vowels and the consonants are each incomplete abstractions, and it is only when they come together that they make a complete sign. In other words, consonants and vowels in Carrier (and in other Canadian Aboriginal Syllabics scripts) are *ion-morphs*. In Carrier, there are additional relationships between consonants that are represented as well. Like the ASL sign for MOTHER, which participates in several lexical families, a Carrier sign like <U> /du/ participates

in several graphical families: that of dV signs, that of Cu signs, and that of CV signs whose C is an alveolar stop.

The scripts of the Canadian Aboriginal Syllabics family have thus far occupied an uneasy typological space between syllabaries and alphabets. Despite the traditional term “syllabics,” they are not syllabaries, as Poser (2003) has carefully explained in the case of Carrier. The values of the consonants and vowels are both represented, and thus Poser states that Carrier is an alphabet. On the other hand, English *Wikipedia*<sup>4</sup> declares the scripts to be abugidas, on the false assumption that the terms *abugida* and *alphasyllabary* are “essentially synonymous” (note 17). In fact, the definitions of *abugida* and *alphasyllabary* are distinct, and they delimit different sets of scripts, as pointed out by Bright (1999) and Gnanadesikan (2017). An abugida is “a writing system in which the basic shapes denote a consonant plus /a/ and the other vowels are designated by attachments to the basic shape” (Daniels, 2018, p. 156). An alphasyllabary “writes each consonant-vowel sequence as a unit... in which the vowel symbol functions as an obligatory *diacritic* to the consonant” (Bright, 1996b, 384, italics in original). In fact, the scripts of the Canadian Aboriginal Syllabics family do not meet either definition. There is no default /a/ vowel, and the vowels are not written with dependent signs that attach to the consonant signs (although the vowels /e/ and /i/ do involve the addition of small lines or a dot). In earlier work I have called this family of scripts “fully vowelised *āksharik* segmentaries” while admitting that “[i]t is an open question ... whether the spatial arrangement of the vowels—in the sense that they overlap in space with the consonants rather than being diacritic on the consonants—is an important typological distinction to make” (Gnanadesikan, 2017, p. 29).

The analogy with signed ion morphs may help clarify the typological place of Carrier and related scripts. As Poser (2003) has stressed, Carrier represents all its consonants and all of its vowels. It is therefore a fully vowelised segmentary, in the terminology of Gnanadesikan (2017). However, typologists have concerned themselves not just with *which* phonological units a script represents but *how* (spatially) they do so. In other words, they have concerned themselves with matters that are, from the point of view of this paper, morphological. Morphologically speaking, an alphasyllabary, or akshara system, is one in which consonants are main signs (stems) and the vowels are affixes. The CAS scripts have ion-morph arrangement, in which there are no affixes but there are simultaneously realized relations between signs. They are therefore not alphasyllabaries/akshara systems but rather their own morphological type, ion-morph segmentaries.

It should not surprise us to find sign-language-like morphology in scripts. Script shares with sign many of the features that Fernald and

---

4. [https://en.wikipedia.org/wiki/Canadian\\_Aboriginal\\_syllabics](https://en.wikipedia.org/wiki/Canadian_Aboriginal_syllabics)

Napoli (2000) adduce to explain the preference for ion-morphs in ASL. Script is produced much more slowly than spoken words. Thus it shares with sign a slower rate of articulation than oral language possesses. Both sign and script are processed visually, where the greater bandwidth allows for more concurrent information than in the auditory channel, offsetting at least somewhat the slower rate of articulation. The benefits, particularly to older learners, of having signs with related meanings have related forms is also relevant to writing, which is also often learned later in childhood or even adulthood. Indeed, when the Carrier script was new, it was known for being easy to learn. Literacy spread quickly from person to person with informal teaching, and “[f]or several decades there appears to have been mass literacy in syllabics” (Poser, 2011, p. 1).

## 6. Conclusions

This paper has revolved around three essential points. First, grammar is amodal. The essentially amodal nature of the human language capacity implies that even secondary communication systems such as writing can be expected to exhibit grammar. Specifically, this paper has examined a number of analogs to morphology that may be found in the structures of the world’s scripts. Among heavily morphographic writing systems, Chinese Hànzì and Maya glyphs have previously received analyses that argue for (in the case of Hànzì) or imply (in the case of Maya glyphs) that composite written signs have morphological structure.

Phonographic scripts also display morphology, and in fact it is important to keep in mind that the morphology of the words of a language (the lexical or L-morphology) is a separate system from the morphology of the graphic signs of a writing system (the graphical or G-morphology). The aksharas of Brahmic scripts are analyzed in this paper as consisting of consonantal main signs (stems) with vowel affixes. In all, inflectional affixes, derivational affixes, stem-triggered allomorphy, levels of affixation, and compounding are identified in akshara-based scripts.

In the Canadian Aboriginal Syllabics (CAS) group of scripts, there are families of CV signs, related in one dimension by shape and in the other dimension by orientation. Generally speaking, the vowel representation (orientation) and the consonant representation (shape) cannot be separated in the way that the consonants and vowel signs can be in Brahmic scripts. These families of signs are analogous to the lexical families of ASL, in which signs for semantically related concepts share most but not all of their manual articulatory parameters. The abstract, partial sign that is shared across a lexical family has been termed an ion-morph. The individual consonants and vowels of CAS may be considered graphical ion-morphs, as in ASL. In the Carrier script, expanded

from the original Cree system, additional features of the signs indicate features such as laryngeal contrasts, adding a further dimension of relatedness between signs and thus further ion-morphs.

The second point is that modality influences form. Comparison between sign languages and spoken languages shows that the expression of grammar is influenced by the modality of the language; the human language faculty is able to adapt to the modality at hand. Written language shares with signed language the feature of being spatially arranged. We should therefore expect to find some grammatical structures in written signed language than like spoken language. The spatial arrangement of writing allows for superfixes and subfixes in the case of the Brahmic scripts and Maya glyph blocks in addition to the prefixes and suffixes (and infixes and circumfixes) found in spoken language. It also allows for the nonconcatenative nature of signs in Canadian Aboriginal Syllabics, analogous to the ion-morphs of sign language.

The third point is that the expression of grammar in a script is independent (at least to some extent) of the language for which the script is used. The overpowering nature of the human instinct for grammar results in grammatical features in scripts (such as inflection in Hànzì or ion-morphs in CAS) that are recognizable from primary (spoken or signed) language but are not found in the specific primary languages for which the scripts are used. Thus, writing systems tap into grammatical faculty directly and not just via the primary language.

This paper is merely one of many (some of which are cited in the Introduction) that have argued in recent years for writing as a modality of language and for the applicability of grammatical analysis to written signs. By applying already-accepted differences in modality to the study of writing, some apparent differences between writing and spoken language have been resolved. It will be interesting to see how far the linguistic analysis of script can take us and what a truly amodal model of grammar would look like.

## References

- Aronoff, Mark and Kirsten Fudeman (2011). *What Is Morphology*. 2nd ed. Chichester: Wiley-Blackwell.
- Baroni, Antonio (2015). "Constraint Interaction and Writing System Typology." In: *Écriture(s) et représentations du langage et des langues. Les dossiers d'HEL*. Ed. by Julie Lefebvre, Jacqueline Léon, and Christian Puech, pp. 296–309.
- Bright, William (1996a). "Kannada and Telugu Writing." In: *The World's Writing Systems*. Ed. by Peter T. Daniels and William Bright. Oxford: Oxford University Press, pp. 413–419.



- (1996b). “The Devanagari Script.” In: *The World’s Writing Systems*. Ed. by Peter T. Daniels and William Bright. Oxford: Oxford University Press, pp. 384–390.
- (1999). “A Matter of Typology. Alphasyllabaries and Abugidas.” In: *Written Language & Literacy* 2.1, pp. 45–55.
- Coe, Michael D. and Mark Van Stone (2005). *Reading the Maya Glyphs*. 2nd ed. New York: Thames & Hudson.
- Daniels, Peter T. (2018). *An Exploration of Writing*. Sheffield: Equinox.
- DeFrancis, John (1989). *Visible Speech. The Diverse Oneness and Writing Systems*. Honolulu: University of Hawai’i Press.
- Evertz, Martin (2018). *Visual Prosody. The Graphematic Foot in English and German*. Berlin: De Gruyter.
- Evertz, Martin and Beatrice Primus (2013). “The Graphematic Foot in English and German.” In: *Writing Systems Research* 5.1, pp. 1–23.
- Fernald, Theodore B. and Donna Jo Napoli (2000). “Exploitation of Morphological Possibilities in Signed Languages. Comparison of American Sign Language with English.” In: *Sign Language & Linguistics* 3, pp. 3–58.
- Fuhrhop, Nanna, Franziska Buchmann, and Kristian Berg (2011). “The Length Hierarchy and the Graphematic Syllable. Evidence from German and English.” In: *Written Language & Literacy* 14.2, pp. 275–292.
- Gnanadesikan, Amalia E. (2017). “Towards a Typology of Phonemic Scripts.” In: *Writing Systems Research* 9.1, pp. 14–35.
- (2021a). “Brahmi’s Children. Variation and Stability in a Script Family.” In: *Written Language & Literacy* 24.2, pp. 303–335.
- (2021b). “S1. The Native Script Effect.” In: *Grapholinguistics in the 21st Century–2020*. Ed. by Yannis Haralambous. Brest: Fluxus Editions, pp. 103–123.
- Johnston, T. (2006). “Sign Language. Morphology.” In: *Encyclopedia of Language and Linguistics*. Ed. by Keith Brown. Elsevier, pp. 324–328.
- Joyce, Terry (2011). “The Significance of the Morphographic Principle for the Classification of Writing Systems.” In: *Writing Systems Research* 14.1, pp. 58–81.
- Klima, Edward S. and Ursula Bellugi (1979). *The Signs of Language*. Cambridge, MA: Harvard University Press.
- Law, Danny and David Stuart (2017). “Classic Mayan An Overview of Language in Ancient Hieroglyphic Script.” In: *The Mayan Languages*. Ed. by Judith Aissen and C. Nora. New York: Routledge, pp. 128–172.
- Liddell, Scott K. and Robert E. Johnson (2011). “American Sign Language. The Phonological Base.” In: *Linguistics of American Sign Language. An Introduction*. Ed. by Clayton Valli, Ceil Lucas, and J. Kristin. Washington DC: Gallaudet University Press, pp. 292–331.
- McCarthy, John J. (1981). “A Prosodic Theory of Nonconcatenative Morphology.” In: *Linguistic Inquiry* 12, pp. 373–418.

- Meletis, Dimitrios (2020). *The Nature of Writing. A Theory of Grapholinguistics*. Vol. 3. Grapholinguistics and Its Applications. Brest: Fluxus Editions.
- Montgomery, John (2002). *How to Read Maya Hieroglyphs*. New York: Hippocrene Books.
- Myers, James (2019). *The Grammar of Chinese Characters. Productive Knowledge of Formal Patterns in an Orthographic System*. London: Routledge.
- (2021). “Levels of Structure Within Chinese Character Constituents.” In: *Grapholinguistics in the 21st Century–2020*. Ed. by Yannis Haralambous. Brest: Fluxus Editions, pp. 645–681.
- Nichols, John D. (1996). “The Cree Syllabary.” In: *The World’s Writing Systems*. Ed. by Peter T. Daniels and William Bright. Oxford: Oxford University Press, pp. 599–611.
- Poser, William J. (2003). “ᑄᑎᑏᑦᑲᑦ. The First Carrier Writing System.” <http://www.billposer.org/Papers/dulkwah.pdf>.
- (2011). “Introduction to the Carrier Syllabics.” <https://www.billposer.org/SyllabicsIntro.pdf>.
- Primus, Beatrice (2004). “A Featural Analysis of the Modern Roman Alphabet.” In: *Written Language & Literacy* 7.2, pp. 235–274.
- Sampson, Geoffrey (2015). *Writing Systems*. 2nd ed. Sheffield: Equinox.
- Snell, Rupert and Simon Weightman (1989). *Hindi*. Chicago: NTC Publishing Group.
- Sproat, Richard (2000). *A Computational Theory of Writing Systems*. Cambridge: Cambridge University Press.
- Steever, Sanford B. (1996). “Tamil Writing.” In: *The World’s Writing Systems*. Ed. by Peter T. Daniels and William Bright. Oxford: Oxford University Press, pp. 426–430.

# The Ideology of “Monographism” and the Advantages of Digraphia. The Case of Lombard

Paolo Coluzzi


*Abstract.* This paper discusses the ideology of monographism and its possible overcoming through digraphia, i.e., the use of two or more writing systems for the same language. After a general introduction, the specific case of Lombard will be discussed as an example. Lombard, a regional language spoken in Northern Italy, is written using different writing systems, more specifically three main ones for the Western variety. As each of these writing systems has advantages and disadvantages, the author sees digraphia as a possible and workable solution, not only for Lombard but also for many other minority or regional languages in the world that find themselves in a similar situation.

## 1. Introduction

Even though many languages still exist in the world that do not have a writing system and are only oral, there is no doubt that in modern times a minority or regional language stands more chances to survive if it can be written down. The fact that the language has a written form can greatly help its status and allows for many strategies of revitalization to be attempted than if it were just oral.

For the same language many writing systems are of course possible; the problem is that sometimes the minority community gets divided over this issue of “graphization” and long and even harsh diatribes have arisen. Believing that one language should have only one writing system can be seen as an ideology of “monographism,” an ideology that is closely related to that of “standard language”. The problem with this ideology is that it is an “either-or” ideology, and it is normally the orthography which gets official support that wins out. The alternative orthography/ies, however, may be around for a long time together with resentment and division, which is not good for language revitalization.

---

Paolo Coluzzi  0000-0003-4571-267X  
Universiti Malaya, Kuala Lumpur, Malaysia, Unit 12-01 Amcorp Serviced Suites, 18  
Jalan Persiaran Barat, 46050 Petaling Jaya, Malaysia  
E-mail: pcoluzzi@yahoo.com

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 67–74. <https://doi.org/10.36824/2022-graf-colu>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

Lombard, an endangered regional language spoken in Northern Italy, is going to be used as a case study in this paper to exemplify the problem with this ideology and its possible solution. After a brief outline on digraphia, this paper is going to discuss the ideology of “monographism,” followed by a general outline of the Lombard used to write Lombard and on how adopting a digraphic or multigraphic system may help the maintenance of the language.

## 2. Digraphia

Digraphia refers to the use of more than one orthography or script to write the same language. For more than two writing systems, the term “multigraphia” could also be used. There are two main types of digraphia: diachronic and synchronic (Dale, 1980). Diachronic digraphia, the most common case, refers to different orthographies or scripts that have succeeded each other over time. Turkish is one example, which was written using Arabic characters until 1928 and is now written using Roman characters. Synchronic digraphia, on the other hand, refers to the contemporary use of two or in some cases more than two orthographies. This may be due to different reasons, mostly religious and political (*ibid.*), but also because the language itself may require different scripts to be written (see for example Japanese) or for didactic reasons (a marginal case according to Dale (*ibid.*)), which is for example the case of Mandarin, which can also be written using pinyin, the official Romanized form. As Dale has explained (1980, p. 12): “The most common type of situation in which a marginal type of digraphia is said to occur is the language-learning situation, or the attempt to communicate something of the sounds of the language to people who don’t know the usual script in which the language is written.”

## 3. The Ideology of “Monographism”

Despite what many people may think, this ideology, which is closely related to the ideology of standard language, has been very strong and pervasive, to the point of having disrupted and hindered quite a few minority language planning efforts. Often, when a writing system is devised by experts or activists for a language that did not have one, some other individuals or groups within the local community or even the academic community may come up with alternative orthographies which are considered better, i.e., more precise, authentic, inclusive or simply more peculiar and distinct from the majority language in the country. For the same language many writing systems are of course possible, some “deeper,” some “shallower,” some using the Latin script, others using

other scripts, such as Cyrillic or Arabic. It largely depends on the purposes or cultural/political orientation of the proposers, and any of these systems has advantages and disadvantages. The problem is that sometimes the minority community gets divided over this issue of “graphization” and long and even harsh diatribes have arisen. I’ve come across a few examples of this ideology and its consequences, such as the opposition between the official orthography for Galician, closer to the Spanish one, and the “reintegrationist” one closer to the Portuguese spelling;<sup>1</sup> or the official orthography for Friulian in Friuli (Italy) and the Faggin system using *baček* diacritics which make it look more Slavic (see Coluzzi (2007) and Coluzzi, Brasca, and Miola (2019)).<sup>2</sup> Something similar is happening in Lombardy, but this will be discussed further on.

#### 4. Lombard

Lombard is one of the languages of the Gallo-Italic group or, perhaps better, of the Gallo-Romance Cisalpine group belonging to the Western Romance family of Indo-European languages, genealogically closer to French and Occitan varieties than to Italian.

According to the 2006 ISTAT survey, about 3.5 million people in the Lombardy region can speak Lombard, i.e., about 36% of the regional population. However, to this figure the speakers of related varieties in bordering areas such as Eastern Piedmont, Canton Ticino and the southern valleys of Chantun Grischun in Switzerland and most areas in Western Trentino should be added. In any case these 3.5 million speakers (and we don’t know how proficient their Lombard may be) are on the decrease—even just by looking at the results of the ISTAT survey carried out only six years before, we can see a decrease of almost 3 percentage points, from 38.6% in 2000 to 35.7% in 2006. We could reasonably deduce that Lombard, in the same way as other Italian regional languages, is losing at least 1/4 of its speakers in every successive generation, which clearly places Lombard among endangered languages.

In fact, according to EGIDS, one of the most well-known scales for the assessment of language vitality, developed by Lewis and Simons in 2010, Lombard like some other Italian regional languages may score, according to the areas, between 6b and 8a. 6b corresponds to the label ‘threatened’, whereas 8a corresponds to the label ‘moribund’. Only two more grades separate the latter grade from the last, 10 ‘extinct’, and this is another clear sign of the predicament Lombard finds itself in.

---

1. For example, “iniciación” and “deseño” would be written respectively as “inici-  
açom” and “desenho” in the reintegrationist system.

2. For example, “cjan” and “palaç” would be written respectively as “čhan” and  
“palač” in the Faggin system.

Even though the total number of speakers is gradually shrinking, a small pool of new speakers is present and very active. For these mostly young speakers Lombard is a second language they have at some point decided to learn, even though chances for using it are not many, particularly in the big cities. In many cases the main domain where they can use the language is the Internet (see Coluzzi (2019)).

The Lombard language can be roughly divided into four main varieties (Sanga, 1997, pp. 255–259; Lurati, 2002, pp. 226–227; Bonfadini, 2010, p. 22):

- Western Lombard (spoken in the provinces of Varese, Como, Lecco, Sondrio, Milan, Monza, Pavia and Lodi, in addition to Novara and Verbania in Piedmont and Canton Ticino in Switzerland);
- Eastern Lombard (spoken in the provinces of Bergamo, Brescia, Northern Cremona and Northern Mantua);
- Alpine Lombard (spoken in the provinces of Sondrio, Trento and Verbania, in Canton Ticino and Canton Grischun in Switzerland);
- the so-called peripheral varieties of the lower lands (spoken in the provinces of Pavia, Lodi, Cremona and Mantua).

So far, each Lombard variety has been written using different orthographies, some more phonetic, some more etymological. For example, the western variety of Lombard, and more specifically Milanese, has been written so far using two main systems (see also Coluzzi (2007; 2008) and Miola (2015): the traditional one, more etymological, and the modern one, more phonetic, used in Switzerland as well. The two systems differ mainly in the way vowels are represented (see Table 1).

| IPA | Traditional              | Modern |
|-----|--------------------------|--------|
| ɔ   | ò                        | o      |
| u   | ó (or 'o' if unstressed) | u      |
| ø   | oeu                      | ö      |
| y   | u                        | ü      |

TABLE 1. The main differences between the traditional Milanese orthography and the modern system as far as vowels are concerned

In both orthographies the consonants are spelt as in Italian, with the addition of the digraph <sg> before <e> and <i> to represent the sound /ʒ/ which does not exist in Italian, and the use of an apostrophe to separate the <s> from <c> and <g> before <e> and <i> so that they are read respectively as /stʃ/ (s'c) and /zɟ/ (s'g), sound combinations that do not exist in Italian.

However, a new writing system was devised by the linguist Lissander Brasca about 14 years ago, and published in 2011, which is currently used

by a dozen activists and “freely” interpreted/adapted by others. The system has been called “Scriver Lombard” and defined as a local-polynomic orthography and its aim is to allow the speakers of all Lombard varieties to write their own local variety in a graphic form which is very similar or even identical to the form in which the speakers of any other Lombard variety would write it, so that the identity and meaning of the words would be easily recognised by speakers of other varieties. This implies that the system cannot reflect directly all the phonetic features of any variety, and the speakers of each variety will need to learn how to write this system that is necessarily the most etymological (deep) and least phonetic (shallow) among the ones used so far.

“Scriver Lombard” looks quite different from the orthographies that have been used so far for the Lombard varieties, which are mostly based on Italian spelling. Whereas the use of vowels is similar to that in the traditional Milanese orthography, consonants are used that are not found in the Italian alphabet, such as <ç>, <j> and <x>, while others are used differently from Italian, such as <q> that can be followed directly by <e> and <i> without the interposition of <u> (corresponding to /ke/ and /ki/), or <g> which is mostly pronounced as /g/ even before <e> and <i>. On the whole, whereas the traditional Milanese orthography is, at least as far as vowels are concerned, a little closer to French and the modern one to German, “Scriver Lombard” is closer to the way Lombard was spelt in medieval literature.

## 5. “Deep” and “Shallow” Systems

As Lüpke has explained (2011, p. 329), “Philologists, linguists and educators have insisted for several centuries that the ideal orthography has a one-to-one correspondence between grapheme and phoneme”. Many lay people who have a limited knowledge of linguistic phenomena also seem to share this viewpoint, including some activists for the local languages. However, even though all these individuals tend to believe that “it is better in an orthography to overspecify than to underspecify, underspecification (or the conflation of several phonemes into one grapheme) can be a powerful tool for the creation of a pandialectal orthography in the case of unstandardized and internally diverse speech varieties” (Lüpke, 2011, p. 332), such as the Lombard local-polynomic orthography.

Shallow systems (traditional and modern orthography) have the great disadvantage that they can only be used in a restricted area, or they need a standardised pronunciation, whereas deep systems, such as Scriver Lombard, are more transparent, flexible and allow for local pronunciations of the language. This means that if on the one hand new speakers may find it difficult to learn how to read and write the advan-

tage will be that they will be able to read and understand all Lombard varieties and a sense of unity of the language will be enhanced. This also means that it will be possible to publish more copies of any written document, from poetry to novels to scientific books, enlarging the audience (any Lombard speaker would be able to read them) and reducing costs. It is because learners seem to be helped by shallow orthographies that reflect the actual pronunciation that linguists such as Sallabank and Marquis (2018, p. 249) have affirmed that “a shallow orthography [...] is easier for beginning readers to process”.

There is a consensus that phonological, in particular phonemic, awareness is beneficial to learning to read, and that shallow orthographies, which make most use of that awareness, are helpful to the learner at an early stage. On the other hand, many, probably most, of the world’s readers use “deep” orthographies where the sound and the letter composition of words are indirectly related or even unrelated. (Sebba, 2007, p. 23)

Returning to the Lombard language, an example of the same sentence in the Milanese variety written using the traditional, the modern and the local-polynomic system can be seen in Table 2.

|                        |  |
|------------------------|--|
| English                | My cousin heard her voice and rushed out to hug her                |
| Italian                | Mio cugino ha sentito la sua voce ed è corso fuori ad abbracciarla |
| Traditional system     | El mè cusin l’ha sentuu la soa vos e l’è cors foeù a brascialla su |
| Modern system          | El mè cūsin l’ha sentüü la sua vus e l’è curs fōö a brasciala sù   |
| Local-polynomic system | El mè cusin l’ha sentud la soa vox e l’è cors fœr a braçar-la su   |

TABLE 2. The same sentence written in the different orthographies

Whereas the last sentence would be read like the two previous ones by a Milanese speaker, it could easily be read by a speaker of Bergamasco, for example, and understood just by knowing that “el cusin” in western Lombard stands for the Bergamasco “ol jerman” meaning “the cousin”. In fact, the same sentence in the Bergamasco variety would be written as: “Ol mè jerman l’ha sentid la so vox e l’è cors for a braçar-la su,” a sentence that is very similar to the one above and perfectly understandable by a Milanese, for instance. The list of frequent words that are completely different in the different varieties is not long and they could all be learnt very quickly.



## 6. Discussion and Conclusions

There are differing opinions on the merits of one or another of these orthographies, but the idea is that in the end only one should be adopted. This is, I believe, an aspect of our dualistic Western culture that fails to realise that adopting more than one system may be the best solution to prevent divisions among activists and speakers. In fact, using all these orthographies in different contexts, i.e., accepting a regime of digraphia, would provide speakers with several advantages.

Some may think that this would be burdensome, but as we have already explained, there are languages in the world that use more than one graphic system. Japanese children, for example, have to learn four different systems at school: hiragana, katakana, kanji (which are used in combination) and even romaji, the Latin script, and this does not seem to be particularly problematic. On the other hand, the Malay language can be written using the Latin or Arabic script, even though the latter is not used much these days. If one system is shallower and one deeper like in China (pinyin and the Chinese characters) or in the Lombard case (the classic or modern orthographic system and “Scriver Lombard”), the shallower system could help speakers (especially new speakers) to learn the local language as the shallow form (pinyin in China and the traditional/modern orthographies in Lombardy) would be closer to pronunciation, whereas the deeper system (Chinese characters in China and “Scriver Lombard” in Lombardy) would allow everybody to enjoy wider communication (respectively with all Chinese speakers and with all Lombard speakers in the region) (see Coluzzi, Brasca, and Miola (2019)).

For the specific case of Lombard, specifically Western Lombard, other advantages can also be seen. Learning the modern system would allow speakers to read comfortably material produced in Switzerland, whereas the traditional system would make reading Milanese literature easy as most of it (mostly poetry and plays) has been written using this older system. Using it would also help not to alienate those older speakers and activists who use and are used to the traditional system.

## References

- Bonfadini, Giovanni (2010). In: *Fécb, cun la o cume fugus. Per Romano Broggin in occasione del suo 85° compleanno, gli amici e allievi milanesi* [For Romano Broggin on the occasion of his 85th birthday, Milanese friends and pupils]. Ed. by Gabriele Iannàccaro, Massimo Vai, and Vittorio Dell’Aquila. Alessandria: Edizioni dell’Orso, pp. 21–33.

- Brasca, Lissander (2011). *Scrìver Lombard, un'ortografia polinomeg-local per la lengua lombarda* [Writing Lombard, a polynomic-local orthography for the Lombard language]. Monça/Monza: Menaresta.
- Coluzzi, Paolo (2007). *Minority language planning and micronationalism in Italy. An analysis of the situation of Friulian, Cimbrian and Western Lombard with reference to Spanish minority languages*. Oxford: Peter Lang.
- (2008). "Language planning for Italian regional languages ("dialects")." In: *Language Problems and Language Planning* 32, pp. 215–236.
- (2019). "New speakers of Lombard." In: *Multilingua. Journal of Cross-Cultural and Interlanguage Communication* 38, pp. 187–212.
- Coluzzi, Paolo, Lissander Brasca, and Emanuele Miola (2019). "Writing systems for Italian regional languages." In: *Journal of Multilingual and Multicultural Development* 40, pp. 491–503.
- Dale, Ian R. H. (1980). "Digraphia." In: *International Journal of the Sociology of Language* 26, pp. 5–13.
- "ISTAT 2006. La lingua italiana, i dialetti e le lingue straniere [Italian language, dialects and foreign languages]" (2006). [www.istat.it/salastampa/comunicati/non\\_calendario/20070420\\_00](http://www.istat.it/salastampa/comunicati/non_calendario/20070420_00).
- Lewis, Paul and Gary Simons (2010). "Assessing endangerment. Expanding Fishman's GIDS." In: *Revue Roumaine de Linguistique/Romanian Review of Linguistics* 2, pp. 103–120.
- Lüpke, Friederike (2011). "Orthography development." In: *Endangered Languages*. Cambridge: Cambridge University Press, pp. 312–336.
- Lurati, Ottavio (2002). "La Lombardia (Lombardy)." In: *I dialetti italiani*. Ed. by M. Cortelazzo et al. Torino: Utet, pp. 226–260.
- Miola, Emanuele (2015). "Chì pòdom tucc scriv come voeurom. Scrivere in lombardo online [Here we can all write as we like. Writing in Lombard online]." In: *Elaborazione ortografica delle varietà non standard. Esperienze spontanee in Italia e all'estero*. Ed. by Silvia Dal Negro, Federica Guerini, and Gabriele Iannàccaro. Bergamo: Bergamo University Press-Sestante edizioni, pp. 79–96.
- Sallabank, Julia and Yan Marquis (2018). "Spelling trouble. Ideologies and practices in Giernesiei/Dgernesiais/Guernesiais/Guernésiais/Djernezié..." In: *Orthography Development for Language Maintenance and Revitalisation*. Cambridge: Cambridge University Press, pp. 235–253.
- Sanga, Glauco (1997). "Lombard." In: *The dialects of Italy*. London/New York: Routledge, pp. 253–259.
- Sebba, Mark (2007). *Spelling and society*. Cambridge: Cambridge University Press.

# Emblematic techniques as textual strategies in non-linear and linear scripts

Liudmila L. Fedorova & Antonio Perri

*Abstract.* Linearity presupposes one-dimensional order in the layout of signs, while emblematic composition refers to a non-linear encoding of information (also linguistic content). Early stages of writing show emblematic arrangement of figurative signs, such the Aztec writing as focused in this paper; while linear order of characters typically emerges with non-figurative units. Yet, since non-linear or emblematic representation of content is primary in textualization practices of any script, the story of writing repeatedly testifies the emergence of multi-linear textual structures, not only in Medieval western manuscripts but also in modern practices of textualization.

## 1. Introduction. Non-Linear Written Texts Between Typologies and Continua

In empirical research on written communication practices, we are often faced with distinctive features of scripts and (more important) of single graphic artifacts which are problematic to fit in standard writing typologies. The latter are, almost systematically, grounded on the structural (i.e., analytic and, ultimately, glottic) principle according to which writing represents abstract language units such as morphemes, syllables or phonemes using *sequential* and (*uni*)*linear* sets of graphic characters as visual signifiers—even though this overall typological pattern can be adequately refined.

---

Liudmila L. Fedorova  0000-0002-2284-6643

Russian State University for the Humanities, Miusskaya square 6, Moscow, 125047 Russia. E-mail: [lfvoux@yandex.ru](mailto:lfvoux@yandex.ru)

Antonio Perri  0000-0002-0453-4849

Università degli Studi di Napoli Suor Orsola Benincasa, Dipartimento di Scienze Umanistiche, Via Santa Caterina da Siena 37, 80132 Naples, Italy. E-mail: [antonio.perri@docenti.unisob.na.it](mailto:antonio.perri@docenti.unisob.na.it)

Authors discussed the whole content of this paper. However, Antonio Perri is the material author of §2, 4 and 5; Liudmila Fedorova is author of §3. Paragraphs 1 and 6 have been written jointly by the authors.

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 75–99. <https://doi.org/10.36824/2022-graf-perr>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

Indeed, it is astonishing that—contrary to what happened for vocal-oral language typology, basically left unchanged for a couple of centuries—as far as writing is concerned dozens of classifications have been suggested by scholars in the last seventy years or so from Diringer and Gelb works—in an endless effort to draw specific (and supposedly discrete, i.e., mutually exclusive) classes, emerging from a small set of unquestionable features.

In order to see why no “arborescent typology” can evade simplification of categories based on unique principles (Klinkenberg and Polis, 2018, p. 93), and before devising an alternative model to deal with internal structure of specific *written texts*—rather than abstract systems or scripts—suffice here to comment on a recent “arborescent” scheme as revised by Fedorova (2021, p. 811).

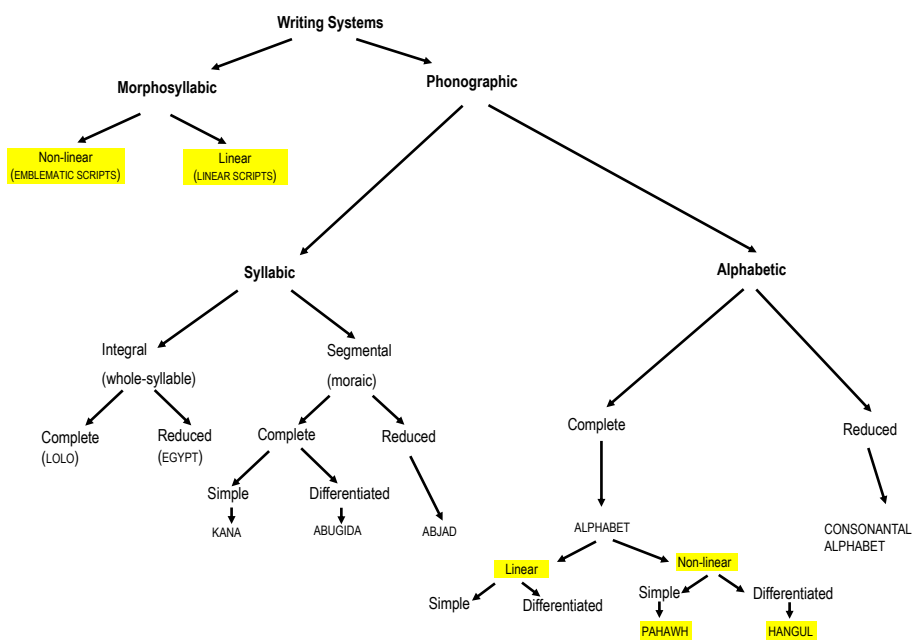


FIGURE 1

In the scheme, another dichotomic feature is added to the usual glottic criterion, this time *not* intrinsically glottic: non-linear *vs* linear arrangement of characters/units and paths of visual ordering. Graphic linearity, in other words, is to be seen as logically independent from *verbal* and *vocal* linearity of reading practices: as Louis Hjelmslev stated over seventy years ago (1973 [1947]), any chain of a linguistic text-as-a-

process is best represented, inasmuch is seen as abstract form, by a horizontal line; however, the latter can be elicited from substantial graphic structures which are differently manifested in visual *medium*. In short, a non-linear arrangement of written units can (and should) always been somehow *linearized* through reading.

By adding this specific feature, therefore, it is possible to include also in the tree (see e.g., the highlighted items, among other scripts, in the scheme above): 1. writing systems traditionally labelled as semasiography, ideography or pictography—aptly and synthetically re-labelled by Fedorova as *emblematic scripts*, and often excluded from the domain of full writing; 2. a couple of phonetic glottic scripts (indeed two alphabetic systems, one recently invented, the other historical) which are unique cases of non-linear patterning in this class (Pahawh Hmong, and Hangul), while we find similar violations also among syllabic abugidas (Fedorova, 2012, 2021).

According to our view, however, when written texts as products and processes of textualization—the planar and stable articulation of units in the visual surface of a written artefact—are concerned, we are dealing with specific procedures of framing which are *unaccountable* through standard typologies of systems-as-codes. This argument, indeed, would not sound so strange, thinking that any theory of translation—and of course, written artifacts are just the endpoint of a special kind of inter-substantial translation process—will not concern anything *but* texts (in their respective roles of source and target). Then, in order to assess the semiotic nature of written text we should dismiss discrete typological models and resort to a couple of interwoven *continua* arranged in a mapping pattern and *not* dependent as such to verbal language—we called them graphic-figurative and graphic-structural (cf. Fig. 2): both are essential features of written textualization procedures of all sorts—since, again, they do *not* concern *writing systems* as such, but *written texts* as a specific subset of visual text.

In the following sections we shall see that *topologically encoded graphic space* in any non-linear arrangement of larger units had a constant role in the story of writing, while the non-linear trend at the level of single characters-units (which has been named *entaxis* by Vaillant, 1999) is less widespread as far as non-iconic or, more generally, non-figurative scripts are concerned; furthermore, it almost disappears in different segmental developments of phonography.

It should be taken into account, from a semiotic point of view, that any iconic representation is (at least) bidimensional; and even that in some pictographic texts there could be a “coded” third dimension: namely, the “sequence” of visual layers or superposed *plateaus* obtained by relative dimensional contrasts on the inscribed surface.

From an evolutive and diachronic perspective, nevertheless, we can assume that the general trend of transition to phonetic writing—

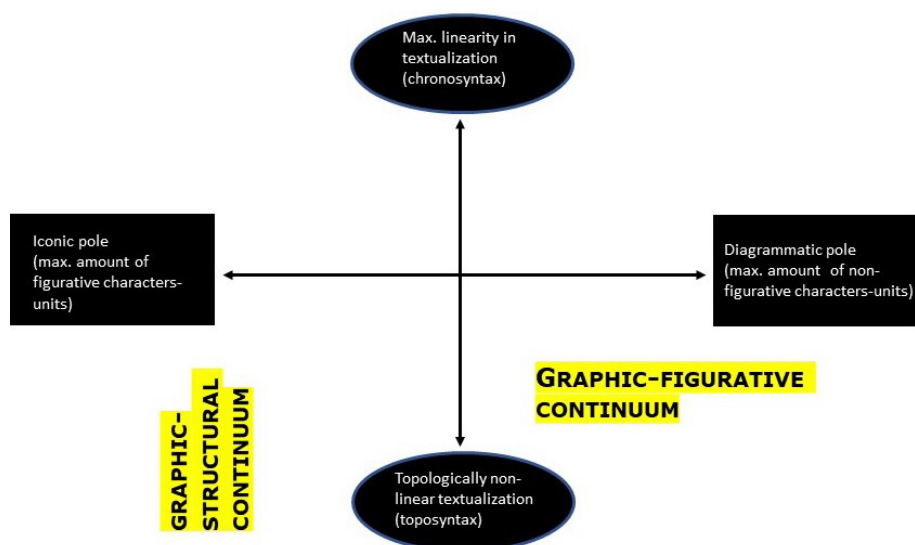


FIGURE 2

together with a widespread generalization of linear paths—did not necessarily meant a total dismissal in the use of pictorial images. Figurative characters continued to be at the basis of developed writing systems for a long period of time (for example in Old Egyptian, or Chinese), and only over centuries in the practice of rapid writing was progressively replaced by what is commonly called a *linear sign* showing a linear (chrono-)syntax, i.e., a sequential-segmental arrangement of written units in the external space of each character—the visual space of written text. A written unit of notation, we could therefore say, becomes *topologically linear* when it acquires a *fixed orientation* in a given sequence, *mapping* the temporality of speech; however, even in Phoenician writing, letters can sometimes unfold in different directions.

A parallel can be seen in the development of writing practice and in the development of speech: as Vygotsky noted many years ago (Vygotsky, 1982 [1934]), there is a stage in the development of intelligence preceding speech, and there is a stage of vocal articulation in the development of speech preceding intelligence; both abilities (mental and vocal) are combined at later stages of ontogenesis in the formation of “visual thinking” (the general technique of visual representation of information), on the one hand, and the development of graphic-figurative forms, on the other, gave rise to linguistic writing. In their combination, the “grammar of pictography” gives way to a more flexible “grammar of language,” which is more capable of expressing through tense, modalities and other categories the dynamics of thinking. Nevertheless, as

far as marginal spheres (vocal abilities and thinking) both remain in the development of language, the sphere of visual two-dimensional representation of information (in diagrams, maps, tables) and the sphere of figurative representation similarly remain in the development of writing.

The purpose of this paper is to sketch in which ways the principle of non-linearity, or “emblematic principle” in two- (or three-)dimensional framing of written space emerge in texts of different eras and cultures; in disparate ways, indeed, those written artefacts violate linear glottic order, but reveal their own methods of organizing information while providing appropriate visual arrangement to coded characters-units.

Section 2 deals with some cases of “internal” non-linearity in the coding of written units of some systems still in use, and with effects such a violation of a classical typographic rule—that of linear typesetting—had for the overall digital project of the Unicode standard.

Sections 3 and 4 are consecrated to a deeper exploration of Aztec writing techniques, from arrangement of glyphs in emblems (§3) to the framed non-linear syntax corresponding to an entire text (§4).

Section 5 will suggest that even in European medieval and modern written tradition we can find brilliant cases of non-linearity in texts, at different layers of structure and complexity—but basically in the arrangements of external space between written units, often organized around figurative or diagrammatically complex principles.

Finally, section 6 will summarize our argument with some concluding remarks.

## 2. Non-Linearity in Writing Units, and *Entaxis*: Cases and Consequences for the Unicode Standard

When non-linear matching between *written units* and speech is consequence of character’s *entaxis*—i.e., of the arrangement of graphic traits internal to a given unit in a writing system (cfr. Vaillant, 1999, pp. 260–61)—we are always dealing with “traditional” but *coded* rules for composing and/or connecting characters of specific scripts. A well-known case is Devanagari, showing special ligatures of signs-characters in cursive script (cf. Fedorova, 2021, p. 819). As Fedorova suggested, in those cases “the reading of components in well-defined order” it is always allowed, and “the enigmatic nature of emblem [i.e., non-linear arrangement] can be perceived only through distorted visual proportion of elements that make reading difficult to non-accustomed readers”. However, those rules are often (but not always, nor completely as we shall see) processable in contemporary digital typographic standards. Then simple, isolated glyphs of Devanagari are coded in Unicode with a single code—as abstract *graphemes*, we would say, if the term could really be

defined cross-graphologically (cf. Meletis, 2022), to which we can raise some doubts; but they change their shape in visualization when connected to others units, according to ligature or relative position in the space around the *akshara* as prescribed—i.e., according to the way letters-characters are joined tighter in script flow. While in Devanagari this “joining” does not always follow a strict glottic or segmental rule—appealing to a prejudicial postulate, indeed, someone would even call these script rules for cursive “(il-)logic”—as we have said before any “graphic” arrangement is uniquely coded in the system; it could, then, be managed by Unicode software, through a display engine. In Fig. 3 (from Cimarosti 2005, p. 92) we can see the “steps” followed by this display engine to give a correct final visualization of the Hindi written word *trimurti* ‘trinity’: the occurrence of the diacritical 094D allows the formation of consonantal nexuses such as <tr> and <rt>, and the last step *re-organizes* glyphs according to the graphic rules of cursive Devanagari writing (with glyph for [i] joined to the left side of syllable it ends as a nucleus: <i-tr> for /tri/, <i-(r)t> for /rti/).

|            |      |      |      |      |      |      |      |      |      |
|------------|------|------|------|------|------|------|------|------|------|
| त          | र    | ि    | म    | र    | त    | ि    |      |      |      |
| 0924       | 094D | 0930 | 093F | 092E | 0942 | 0930 | 094D | 0924 | 093F |
| त          |      | ि    | म    |      |      | र    | त    | ि    |      |
| t          |      | +r   | +i   | m    | +ū   | r+   | t    | +i   |      |
|            |      | त्र  | ि    | म    |      |      | र    | त    | ि    |
|            |      | tr   | +i   | m    | +ū   | r+   | t    | +i   |      |
|            |      | ि    | त्र  | म    |      |      | ि    | त    |      |
|            |      | +i   | tr   | m    | +ū   |      | +i   | t    | r+   |
| त्रिमूर्ति |      |      |      |      |      |      |      |      |      |

FIGURE 3

However, when rules for composing syllabic blocks are not intuitively grasped by a writer/user nor formalizable through specific algorithms—as it happens for Korean Hangeul—the Unicode consortium solves the puzzling situation in a paradoxical way (cf. Perri, 2007). In the Unicode Standard 14.0 (Unicode Consortium, 2021, p. 141):

The Unicode Standard contains both a large set of precomposed modern Hangeul syllables and a set of conjoining Hangeul jamo, which can be used to encode archaic Korean syllable blocks as well as modern Korean syllable blocks. This section describes how to – Determine the canonical decomposition of precomposed Hangeul syllables. – Compose jamo characters into



precomposed Hangul syllables. – Algorithmically determine the names of precomposed Hangul syllables.

The paradox we alluded to is the “double coding” of each syllabic block of Hangul, e.g., the simple syllable /a/ displayed in Fig. 4—both as a precomposed basic unit, or as competent intentional composition, by an expert user, of two individual jamo characters. However, the *quantitative effect* in terms of number of glyphs-units coded is astonishing, since we are dealing with a “special” alphabet. Modern Korean allows the occurrence of 19 consonants as syllable onset, 21 vocals and 27 consonants as a coda, thus combining in 399 blocks of two characters and 10.773 of three characters each: thus, the total amount of 11.172 syllables (a figure confirmed by modern grammars) has its specific code in the Unicode standard.

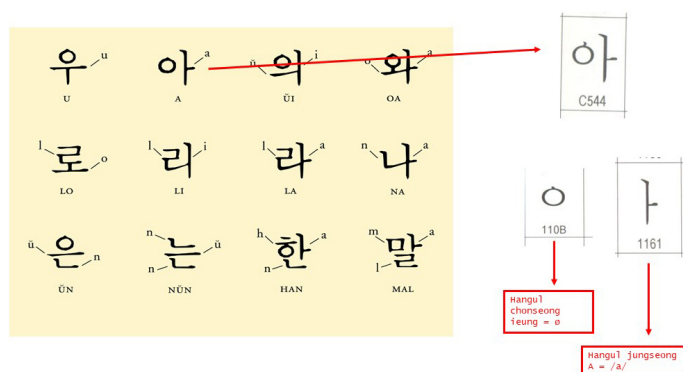


FIGURE 4

### 3. Aztec Linguistic Emblems: Arrangement of Glyphs

First, we will briefly introduce Aztec writing, together with the logic to write place-names and proper names of historical individuals—the only parts of this script so far acknowledged as writing in the traditional or glottic sense (cf. also Fedorova, 2009).

The readable composition of pictorial glyphs in the function of such nominations is called *linguistic emblem* (term introduced in Fedorova 2009). The linguistic emblem usually combines two or three signs, the order of reading is not determined, and the image can vary a little due to ingenuity of the writer, so that some glyphs have no phonetic correspondence in the nomination, while some phonetic fragments (mostly suffixes) have no visual representation. The “text” of a pictorial codex can

be interpreted or reconstructed according to certain rules of arrangement of components, symbolic or linguistic emblems, in the “written” space, for different genres (e.g., in case of a narrative, the emblem of main(s) event(s) placed in the centre of textual space, while marks of years at the margin). Such frame disposition, corresponding to the content of an oral text, can be treated as a *textogram*—to use Dyakonov’s term (1976).

A *textogram* combines then the iconicity of images and their positional function in a pluri-dimensional but meaningful space differently structured by a *toposyntax*, in the sense of Klinkenberg and Polis (2018, pp. 65–66), according to whom it «makes use of spatial dimensions» so that «values of order and succession make way for values of simultaneity».

Therefore the pictorial textogram includes symbolic and linguistic emblems: the former having only pictorial value, the latter understandable as readable glyphic compositions. Symbolic emblems can have nevertheless correspondence to the language units specific for the Aztec culture: to so called “binoms” (Spanish nahuatl scholars named them *difrasismos*), e.g., an emblem of WAR—arm and arrows—corresponds to the nahuatl binom *in mitl in chimalli* ‘the arrows, the shield’. Another example is a symbolic emblem of conquered city—falling and burning temple corresponds in nahuatl to a more elaborated formula: *in teocalli popoca, tlatla iicampa, in montemayau, tepebualiztli*, ‘the temple emits many smoke, his back is on fire, it is the conquest, the defeat’ (cfr. Perri 1994, pp. 166–67).

Yet, for a native reader, there could be no sensible difference in comprehension regarding both types of emblems.

The task of reading those linguistic emblems in textograms involves the ability—something similar to “intuition” alluded to by Elkins, but well known to indigenous readers—to *switch* and to *transduce* from visual composition to a phonetic pattern, since there are no specific graphic markers guiding this switching and the units of both codes do not basically differ visually (as we are used to think every time we are confronted with the Western text-image contrast). However, there are well-known graphic (and functional) positions in which genuinely phonetic signs can occur: this is the case for emblems occurring besides pictorial images of humans (they designate a proper name or ethnonym), or besides glyphs of settlements or conquered cities (in this case, they are place names). Still, due to its pictorial nature and to its multiple linguistic values (Whittaker 2021, pp. 54–55), not every emblem of name is purely phonetic; logo- and morpho- are also linguistic units, therefore they are readable as these words or roots.

Yet, according to many scholars, we can prove a strictly phonetic reading of an emblem only when it resorts to a rebus device in representing some components (homonym’s images).

Regardless of the linguistic level expressed by glyphs (words, morphemes, syllables), we can conclude that Aztec linguistic emblem is always a complex graphic nomination, phrase provided with *unambiguous reading*.

Usually, it is a composition of two or three components-units, corresponding to a compound name or sometimes representing its “interpretation” by the writer (Aztec names often show the linguistic structure [(root+root)+suffix]). However, suffixes often are not expressed with a specific graphic mark: it is assumed that the whole word is understood (and read) even without the ending marker. In some cases, however, the locative suffix is expressed by the very topologic arrangement of glyphic components. When the written linguistic emblem of a personal name or the name of a tribe is expressed, however, it is usually attached behind or above the character’s head with a thread-line, forming a kind of *graphic ligature* (called by Galarza e Maldonado Rojas, 1986, pp. 145-146 *lazo gráfico*).

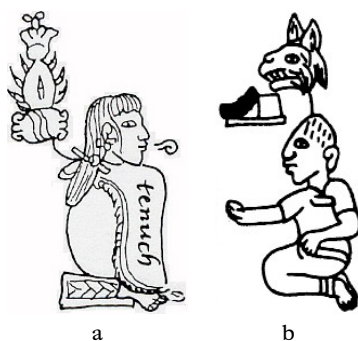


FIGURE 5

This is the case for *Tenoch* in Fig. 5a (from Codex Mendoza, f. 2r), where the emblem transcribes the personal name of the Mexica priest-ruler decomposed by a scribe: *te-tl* ‘stone’ + *noch-tli* ‘cactus’; while in the next example (Fig. 5b), a place-name glossed by the Spanish interpreter *Coyocac*, the textogram is possibly to be understood as a tribe name (*coyo-tl* ‘coyote’ + *cac-tli* ‘sandals’, phonetically expressing the non ethimogizable name *Coyuca*), ethnic label of the female person in the toponym *Coyucac*, ‘in the place of Coyuca (of the women from Coyucac)’ (cfr. Peñafiel, 1885, pp. 84-85; 1895, p. 69, Codex Mendoza ff. 2r e 13r).

The ways of labeling with a place-name a settlement symbol can vary, writing sometimes a complex ‘name of settlement’—as we have seen for the emblem in Fig. 5b.

Locative suffixes in the place-names can be represented:

1) with glyphs-homonyms (Figs. 6a and 6b):

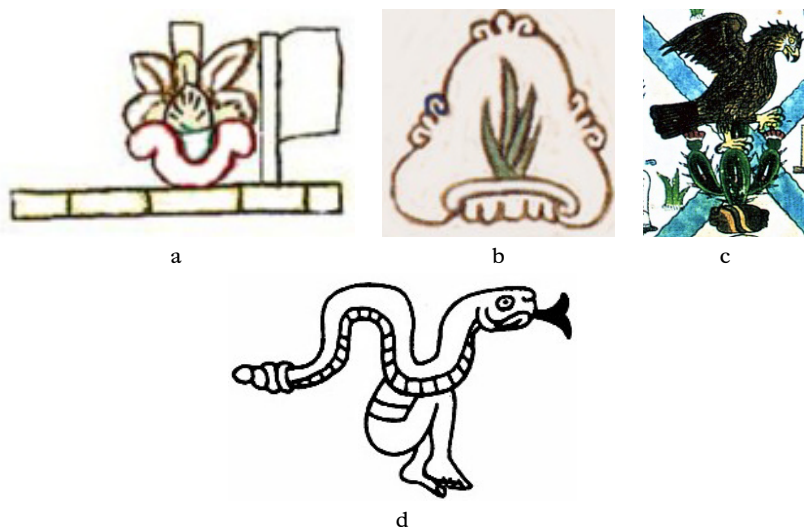


FIGURE 6

ACAPAN (*aca-pan*, ‘on the reeds’) with a glyph FLAG (in nahuatl *panthli*) for *-pan* ‘on’ (Fig. 6a); ACATLAN (*aca-tlan*, ‘among the reeds’) with a glyph TEETH (in nahuatl *tlantli*) for *-tlan* ‘among’ (Fig. 6b);

2) with the mutual arrangement of glyphic images (Figs. 6a–d)

In the examples of Figs. 6a and 6b the locative meanings are also presupposed by the position of glyph expressing the root morpheme: in *Acapan* the reed (*acatl*) in a sort of pot is placed *on* a platform; in *Acatlan* the same reed (*acatl*), this time represented as a simple plant with its leaves, is placed in the middle of glyph HILL (*tepetl*, non-readable base for settlement), thus standing for ‘among’.

The suffix *-titlan*, ‘among’, ‘where there is a lot of ...’ in (Fig. 6c) TENOCHTITLAN (*te-noch-titlan*) is rendered with a glyph of EAGLE (non-readable symbol with mythological reference) “sitting down” *among* the cactus branches (*nochtli*) on a stone (*tetl*). Yet locative suffixes are mostly not represented at all, or can be only arbitrary supposed, as the place-name in Fig. 6d COATZINCO (*coa-tzin-co*): in the emblem, locative suffix *-co* is not represented phonetically, yet it is given by a hint since the glyphic image for *-tzin* (diminutive suffix, homonym to ‘the man’s lower half’) is brought by a snake—like a burden *on* the snake.

The components-units of a linguistic emblem can also be tied together by using different graphic techniques of plastic combining (Galarza named the latter *lazos plasticos*, cf. Galarza, Maldonado Rojas,

1986, p. 146): we can call those techniques, respectively, as *juxtaposition*, *addition*, *incorporation*, *syncretism*.

*Juxtaposition* is a free combination of glyphic images co-located (maybe at some distance) usually without any semantic relation in the extralinguistic world: the arrangement, then, forms a whole composition that can be interpreted in “logical” or fantastic way (as a graphic game). We can speak of *addition* in any case of “logical”—from an iconic and extralinguistic point of view—combination of two (or more) images, while *incorporation* is to be intended as insertion of a relatively small image-unit in a “bigger” one. Finally, *syncretism* is a special case in which two pictographic units cannot be isolated in the whole—as a sort of blending (in terms of compounding in linguistics).

The Figs. 5a and 5b above (the same can be said for Fig. 6c) are examples of two of those abovementioned techniques: in Fig. 5a we have a case of *graphic addition*—since a cactus on a stone is a referentially meaningful and complex image linguistically readable; while in the ethnonym of Fig. 5b there is a mere *juxtaposition*, since the units of coyote and sandal do not form a coherent “image” in any way.

Figs. 6a and 6d also are examples of juxtaposition, while Fig. 6b presents a *double* incorporation of REED and TEETH in the glyphic form of an (unread) HILL.

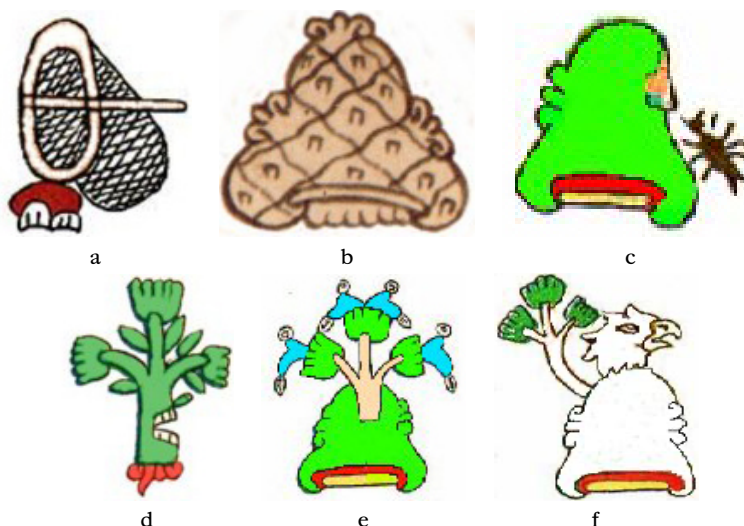


FIGURE 7

Furthermore Fig. 7a (a place-name glossed *Matatlan*, ‘in the net’) from Codex Mendoza (f. 10v) is also considered as juxtaposition of two

unrelated components: *mata-tl* ‘net’ + *tlan-tli* ‘teeth’—occurring in this case as phonetic transcription of the nahuatl suffix *-tlan*, ‘among’. The glyph occurring to write down for the same place-name in the Historia Tolteca Chichimeca (33r), and reproduced in Fig. 7b—therefore an iconic-graphic variant of the former—shows on the contrary a “fantastic” agglutinative image: indeed, it is a sophisticated case of an emblem-nomination combining *incorporation* and *syncretism*: first, the glyph-unit for TEETH is inserted as special graphic variant into the contour of a cave at the foot of a HILL (*tepetl*), to be understood as general glyphic mark for any inhabited settlement (*altepetl*, a sign which can also bypass an explicit reading); second, the mountain or hill is somehow “enclosed” by a net—and this is indeed a syncretic, graphically unsegmented expression of the whole.

The place-name glossed *Yaca-pich-tlan* (or *Yaca-pitz-tlan*), shown in Fig. 7c and still from the Mendoza (ff. 8r and 24v) is a telling example of the techniques of *incorporation*, *syncretism* and *addition*: the name structure could be interpreted as ‘place where there are many sharp objects’ (maybe stones or thorns?; or, somehow more literally, sharp noses?). But to write it the Aztec painter resorts to a HILL—again as a generic visual marker-symbol of site, village—showing a NOSE to its side (*syncretism*, the two units are merged), together with a sort of BUG (the linguistic value of which, however is not undisputed); the latter is simply *added* at the right of the glyph, below the NOSE, and as if it may bite it. Then *yaca-tl*, ‘nose’, *pitz-* as a root of words *pitzabuac* ‘thin’, from the verb *pitzabua* ‘get thin’, or *pitzcua* ‘to pinch’, or *pitztli* ‘fruit stone’ (Wimmer 2006, Siméon 1885, cf. Penafiel (1885, p. 247; 1895, p. 321) together convey *yacapitz(-abuac)*—‘sharp’ (perhaps the Aztec word for ‘sharp’ is somehow semantically related to the roots cited, but this is not obvious); ‘sharp’ (perhaps the Aztec word for ‘sharp’ is somehow semantically related to the roots cited, but this is not obvious).

Furthermore, the place-name emblem (Codex Mendoza, f. 39r, Fig. 7d) glossed *Abuacatla(n)*, ‘where there are many avocado trees’ is an instance of *incorporation* creating a “fantastic” image of a ‘tree’, with ‘teeth’ writing down the suffix *-tlan*. The specific kind of tree is not marked (maybe the suffix indication was enough). But Fig. 7e glossed *Abuatepec* (*Abua-tepe-c* ‘on the oak hill’) gives a hint to the name of TREE—*abuatl* ‘oak’ by incorporating glyphs of WATER *a-tl* in the crown of the TREE as phonetic complement *a-*. Fig. 7f shows another place-name from the Mendoza (f. 5v) glossed *Cuabuabcan*, ‘the place of possessors of eagles’ (in the sense ‘the place of eagles’); it exemplifies the *addition* of a tree and an eagle’s head to the glyph of HILL. The tree (*cuabu-itl*) this time is to be read as a (redundant) phonetic complement for homonym *cuāub-tli* ‘eagle’; the both are in juxtaposition to each other; the possessive suffix *-buâ-* and locative *-can* are not directly represented.

So, as we tried to show, different graphic devices are used to create a fixed composition, with necessary and sufficient components to make the emblem recognizable and distinctive from other phonetically close place-names.

#### 4. From Entaxis of Emblems to Toposyntax of Textograms

While all Aztec emblems are non-linear and therefore bi-dimensional items, an “ideal” laying out of glyphic components in linear order (seen as an attempt of reading an emblem through the practices of an European reader who tries to decompose the unit in analyzing it) is sometimes possible: see the case of place-name from f. 20r of the Mendoza glossed *Tepetlacalco*, Fig. 8 [((*te-petla*)-*cal*)-*co*] ‘house in a cage of stone woven net’ = ‘stone cage’. The glyph for *tetl*, STONE is given unfold above and beneath as double addition, including a house in woven net (syncretic image). In this visual arrangement one can see a resemblance to linguistic morphological technique of *circumfix*, an affix “surrounding” a root (in this case, a blending compound).

But of course, in the more sophisticated cases (such as the toponym glossed *Xalatlaubco*, Fig. 9: *xal-atlaub-co* ‘in the sandy canyon’, also occurring in Codex Mendoza f. 10r) only through bidimensional entaxis it is possible to express the occurrence, in reading, of a word such as *atlaubtli* ‘canyon’: it is, indeed, an invisible canyon—represented as “empty space” between two mountains and thus expressing a pictorial, in principle unreadable image. Yet, the painter-writer explicitly marked in the emblem also readable linguistic formants, using in the segmentation the glyphs WATER (*a-tl*) as phonetic complement of the nahuatl word *a-tlaubtli*, and then adding the derivative morphogram SAND (*xal-li*) as a prefix. This syncretic image is non-linearly written by the readable combination of two glyphs, as a “logic” tie of water on the sand.

In order to contrast such non-linear patterns or *entaxis* of emblems with the unilinear syntax of segmental scripts, we can call this kind of framed space, when external to single emblems—which in many ancient scripts, such as predynastic Egyptian, or Sumerian tablets from Uruk (cf. Fig. 10) is displayed mainly at the macrolevel of layout or *textogram*—*synsemia* (Perondi, 2012) or *toposyntax* (Klinkenberg and Polis, 2018).

Fig. 11 shows the patterning of an emblematic-*synsemic* space linguistically framed in (half a) page from the second part of Codex Mendoza (f. 20r), the so-called *matricula de tributos*. Actually, to account for the whole section of this tribute register we should also consider the verso of the folio, since the post-Conquest copying from ancient Mexican *tira* (long strips of inscribed text made by *amate* paper or deer skin, similar to classical Mediterranean *volumina*) to European paper sheets caused a rearrangement of pictographic layout which at times obscured the clearcut

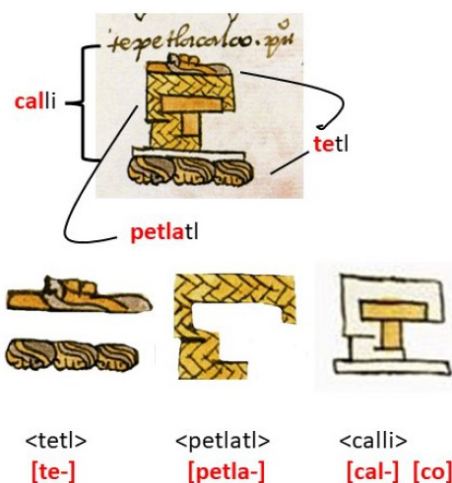


FIGURE 8

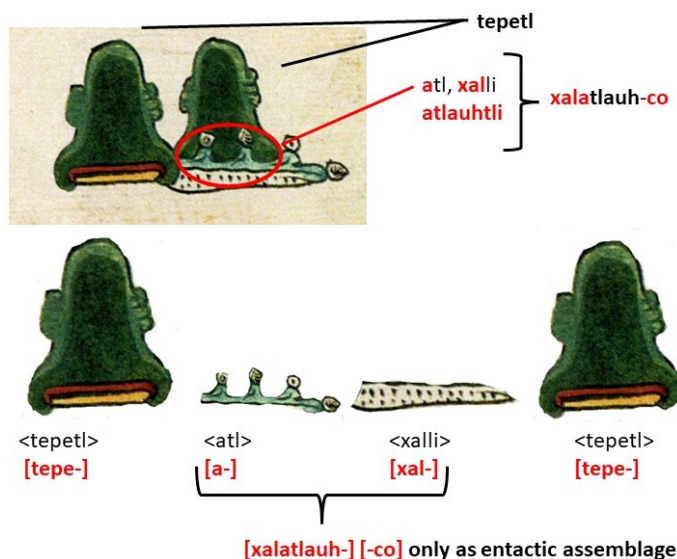


FIGURE 9

pattern of the original (for a hypothetical reconstruction of the original frame, in the horizontal form of the *tira*, see Perri, 1994, pp. 176 ff and 2001, pp. 10-13). For the sake of simplicity, however, we will limit here our analysis to the recto.



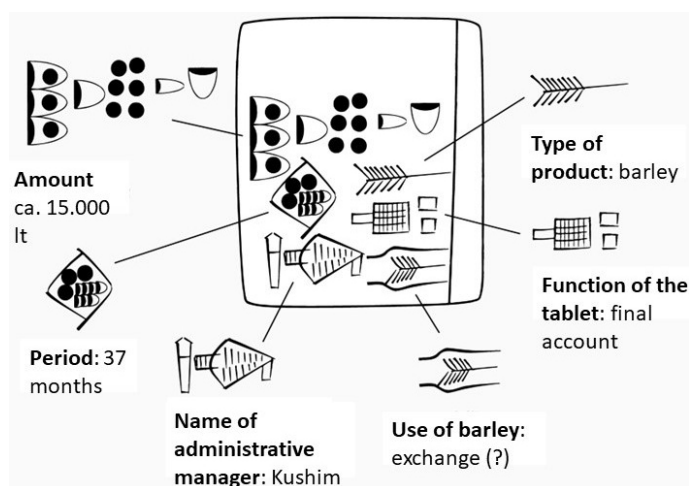


FIGURE 10

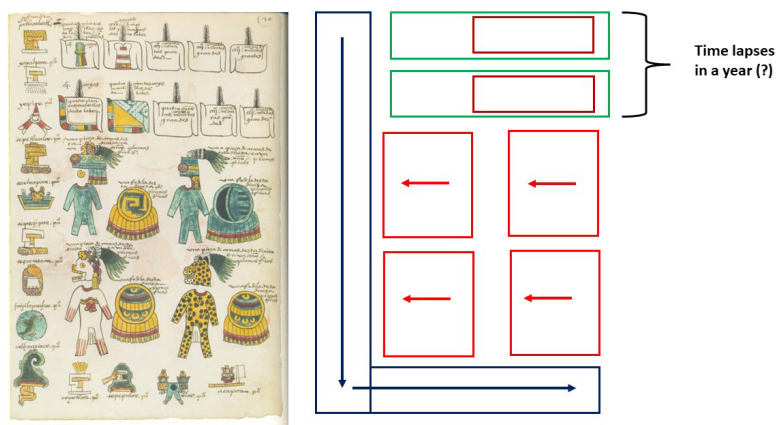


FIGURE 11

Looking at the Fig. 11 (and comparing this page with other sections in the Codex Mendoza) there emerges a “graphic template” or “genre” in which a coded space articulates linguistic (in this case economic) content in ways fostering unambiguous and well-established readings (Perondi, Perri, 2018, pp. 42-44). We are confronted to a pattern where the external frame (the blue box in Fig. 11) is a written list of glyphic place-names—i.e., villages giving the annual tribute to Mexican rulers; while in the internal space we find different categories of pictorially represented items required by Aztecs, ordered according to a definite se-

quence: loads of cloaks, thongs and female dresses first, in ranks at the top of the written space (see the green boxes in Fig. 11); then precious warrior costumes (red boxes). All those tributes, when their pictorial form allows for orientation—as in the case of warrior costumes hats—are oriented to the left, thus “closing” the section which is read from left to right starting from external place-names frame.

It is significant, moreover, that for some items the single tribute is repeated more times with the same quantity (in this case the numeral for ‘400’, *centzontli*, a glyph representing a lock of hair): the loads of white cloaks to be paid as tribute (*centzontli iztactilmatli in tlamalmalli*), indeed, are drawn six times. The only reason to account for this “multiplication” in two ranks of same glyphs—considering that it would surely be possible to write down the tribute in a unique account just linking a single pictorial glyph to the total quantity numerically expressed, as we can see in other sections of the same Mendoza (cfr. Fig. 12, taken from f. 26r and depicting a tribute of 4.000 bulrushes seats, *petlaicpalli*, and the same amount of mats, *metatl*)—is to assume that the notational strategy is aimed at transcribing a definite order in time lapses for payments of this kind of tribute during a single year (which is the period covered by the register, cfr. Perri, 1994, pp. 190-194).

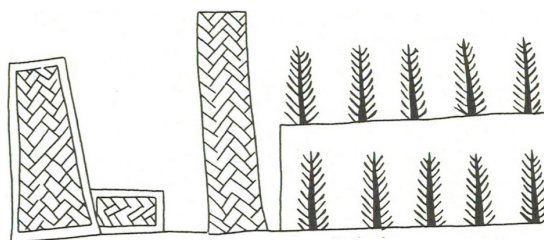


FIGURE 12

The purple frame internal to the green one in Fig. 11, then, is supposed to highlight the fact that one-year-tribute in white cloaks loads was in fact delivered in six times, i.e., every three months in the civil traditional calendar of eighty months of twenty days traditionally in use before Conquest. The theoretical question which arises can be summarized as follows: are those non-linear visual features conveyed by topological distribution of information merely connotative superstructures of a written text which has to be basically read/decoded as a linearized and syntactically coherent flow of speech? In other words, should the fact that we are not told in a written chunk of pictorial text the effective period of delivery for each load of cloaks—this is what happens, as far as we know, in Sumerian tablet of Fig. 10—mean that this is *not* ex-

pressed *at all* nor linguistically retrievable? The answer, still, is “no”: the written text as framed in the genre-specific pattern does not need to be somehow re-transposed in a precise linguistic sequence to be correctly understood (Perondi, Perri, 2018, p. 46).

## 5. Figurative Entaxis, and Emblematic Non-Linear Framing of Written Space in European Tradition



FIGURE 13

Since non-linear internal assemblage of units is typical of figurative or predominantly iconic scripts, it is of no surprise that we find traces of *entactic* non-linearity in some modern notational system never regarded as writing, such as European coat of arms. As Mounin stated fifty years ago (1970), the coding of this restricted notational system is not linguistic in the sense it should be regularly connected to a phonetic reading; however, at times there are congruent linguistic keys: in Fig. 13 we see the coat of an Italian nobiliary family, whose name is Bracci, and the ‘arm’ *emblem* is logographically transcribing this reading.

In any case, since according to Mounin discrete units in coats of arms are never linear, because they are connected to a global reading in a space *not imposing any preferential order*, they are indeed *emblems* but not *proper* writing—therefore not linguistic emblems as in the case of Aztec place-names; however, they are undoubtedly conventionally coded and perform a function similar to that of proper writing.

In order to find instances of synsemic written space in modern alphabetic European tradition, then, we have to approach special kinds of written text where specific expressive needs impose a *topologically encoded graphic space* to suggest multiple or alternative paths of reading.

As a telling example, we can look at the double page of the codex of Guilielmus Peraldus *Summa de virtutibus et vitiis* (half of XIII century, cf. Fig. 14). Peraldo’s text is indeed a huge classification of vices, with a large array of examples (quoted from Bible and other texts). In other words, it is an inventory. In the double page, then, as Lina Bolzoni said every *locus* of the picture is inscribed, but in order to understand what we see, we must not only read the inscription but draw our attention to the *place* in which has been situated. We have to slowly retrace, step after step, the compositive plot [...]: the Christian life as a fighting knight, with his virtues as weapons (2002, pp. 62 ff). The title in Latin at the top of right page, *Man’s life on the earth is soldiering*, is a sort of key to the correct interpretation. Bolzoni showed that the *loci* in the



FIGURE 14

picture reproduce the structure of a dialectical contrast between, in the left page, the seven capital sins (with minor vices deriving from each) and, confronting them on opposite orientation, the seven gifts of Holy Spirit represented by the icons of the doves and the seven beatitudes of the Sermon on the mountain, in the seven phylacteries held by the angel. Synsemic space topologically articulates a number of mental images, for meditation, knowledge, memory, introducing a flexible and “open” relation between writing and pictures.

In this case, of course, mnemonic function of images is quite clear: the writer used a coded space to frame argumentative sequences, but this time emblematic pictures must be thought as visual support for a literate preacher, often addressing illiterate listeners.

Another interesting case is the *Turris sapientiae*, reproduced in Fig. 15 from a printed woodcut, ca. 1475. The title, at the bottom of the page, *Turris sapientie legatur ab inferiori asce[n]de[n]do p[er] seriem l[itte]raru[m] alphabeti* gives to the reader specific instructions: he is supposed to proceed from bottom to top, ascending the iconic structure by following the order of letters at the left margins.





FIGURE 15

Elements of the figurative building analogically express, through their mutual locations and relations, the relationships in the field of knowledge of the *Wisdom* (i.e., the true knowledge inspired by God). We have a multiplication of *loci*, thus *synoptically seeing virtues and their components* via structured correspondences in a coded order. As wrote Antinucci, “the ‘physical’ form of the tower as it is represented *makes we see* what is the relation between concepts linguistically [i.e., alphabetically] expressed in the text [with a non-linear framed space]: it is radically different from linear order, where this relation should be mentally inferred and pieced together”. The *Turris* is significantly named “*Speculum theologiae*, i.e., ‘mirror’, visual representation of theology” (2011, p. 122).

Perhaps the most astonishing example of multidimensional structuring of written space through emblems, however, is the famous *Liber figurarum* by the mystic from Southern Italy Gioacchino da Fiore (manuscript written and illuminated at the beginning of XIV century). In Fig. 16 we reproduce the subtlest and cultured page of Trinitary circles.

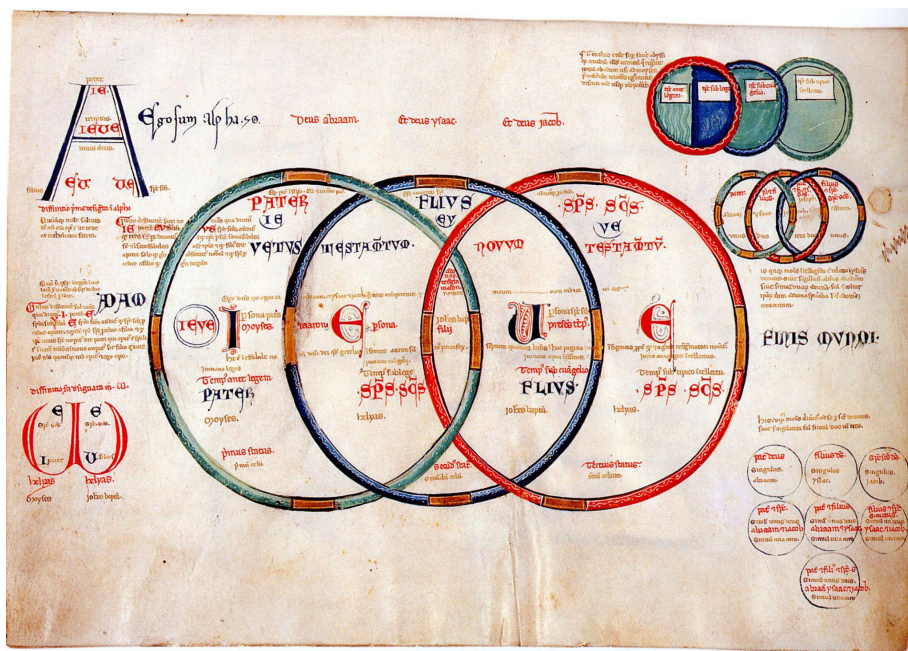


FIGURE 16

In this case the analogically framed space illustrates the relationships and correspondences between the Three Persons of Trinity, and between trinitarian Persons with other doctrine elements visually rep-

resented as theoretically articulated diagrams, such as the Alpha-Omega tetragram (to the left), human history (from left to right, following the intertwined rings at the centre), and some biblical *figurae* (Perondi, 2012, pp. 178-183). The interplay of multiple analogical relations, iconically transposed through diagrams in which written symbols and non-linear logical relations are inextricably connected, condenses a complex doctrinal argument verifying—more than five centuries early—the famous Peircean statement about diagrams: “Diagrammatic reasoning is the only really fertile reasoning” (CP 4.571).

It is useful to note that contemporary examples of emblematic non-linear space in a typographic European domain occur often in visual poetry—cf. e.g., the Stephen Themerson’s (1949) English translation of a Chinese poem by Li Bo, VIII century A.D. reproduced in Fig. 17. In this poem, the use of unusual typographic devices such as *internal vertical justification* increase the topological patterns of reading, and transpose in alphabetic relations the “visual rhymes” occurring in original Chinese text in logographic characters. Looking at this case, we cannot but refer to a suggestion by Anne-Marie Christin (2009<sup>2</sup>, p. 17): according to the author, indeed, space is the only formal feature identical in picture and writing, but her statement is wrong if the visual space is conceived in the form of a screen, something abstract and “empty”. On the contrary, it is always a *coded textual space* which informs any interpretation of a visual artifact, providing an integration with diverse and specific interpretive practices (seeing *vs* reading) often overlapping and mixing.

## 6. Concluding Remarks

We can finally revert to the quadrants of mapping pattern of Fig. 2, since an evaluation of specific *written texts* (and others not mentioned in this paper) allows to fill each of the quadrants in the articulated domain formed by the two visual continua *not* glottic-dependent (cf. Fig. 18).

In this paper, while recognising the importance of traditional glottic typologies, we focused also on their limits:

- first, if, as Gelb suggested many years ago (“there are no pure systems of writing just as there are no pure races in anthropology” [1963, p. 199]), neither there are pure and coherent and no pure and coherent labels to describe all divergencies and specificities: then, for example, Aztec writing could be seen *at the same time* as pictorial, logographic, morphophonemic and phonetic—depending on the specific “genre” of written text, the period, the kind of linguistic content and so on;
- second, perhaps more important, the glottic criterion is not enough, as such, to deal with a systematic analysis of textual products or artefacts;



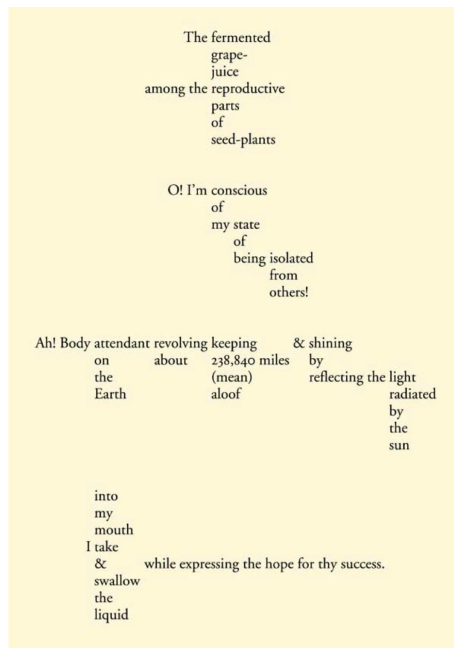


FIGURE 17

- third, the ambivalent status of linearity in writing do not find an adequate treatment within the *visible speech* approach.

According to our view indeed linearity, in any writing system, is a semiotic prerequisite in order to assure an actual matching between visual-graphic expressions or units and (sequential, temporarily “linear” in Saussurean terms) bits of speech; however, while a sequential ordering of reading is always to be assumed, many systems of visual (and coded) graphic signs exploit the (at least) bidimensional visual space both (a) to form/construe written characters-units by joining minimal traits in non-linear paths (entaxis of the internal space); and (b) to articulate written texts combining those units in *non-linear visual layouts* or external toposyntactic space (significantly framing the written space in view of a correct and complete reading-understanding of linguistic content).

Moreover, while linear paths in writing emerge with non-figurative images and, more systematically, with (a more or less) complete phoneticism, the story of writing repeatedly testifies the use of multi-linear structuring patterns.

This fact, in the end, is clearly explained—adopting the integrational approach to writing fostered by Roy Harris (1986, 2000)—when we con-



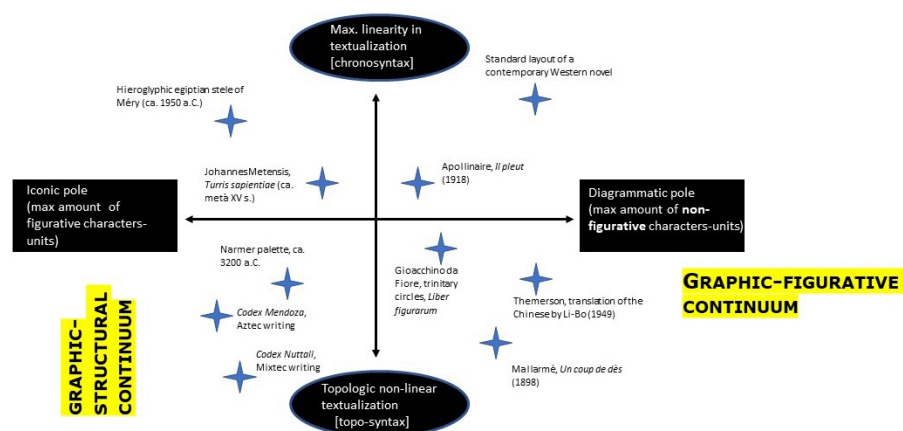


FIGURE 18

sider the semiotic nature of any *objectual space* inscribed, since “in textualized artefacts the text always functions as a sign in ways which are not exhaustively described by giving a merely ‘linguistic’ account of what the text say, of the linguistic [written] forms used, and of their [glottic] meanings as contrasted with other forms available in the [spoken or written] language” (Harris 1990, p. 217).

## References

- Antinucci, Francesco (2011). *Parola e immagine. Storia di due tecnologie*. Rome-Bari: Laterza.
- 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.
- Bolzoni, Lina (2002). *La rete delle immagini*. Turin: Einaudi.
- Christin, Anne-Marie (2009). *L’image écrite, ou la déraison graphique*. 2nd ed. Paris: Flammarion.
- Cimarosti, Marco (2003). “Dodici anni di Unicode.” In: *Progetto grafico* 1, pp. 84–95.
- Consortium, Unicode (2021). *The Unicode Standard. Version 14.0—Core Specifications*. Mountain View, CA: The Unicode Consortium.
- Дуаконов, Игор М. [Дьяконов, Игорь М.] (1976). “Протошумерские иероглифы [Proto-Sumerian hieroglyphs].” In: *Тайны древних письмен: Проблемы дешифровки [The mysteries of ancient scripts. Problems of de-*

- ciphering*]. Ed. by Igor M. Dyakonov [Дьяконов, Игорь М.] Москва [Moscow]: Прогресс [Progress], pp. 569–571.
- 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. 258–275.
- (2012). “The development of structural characteristics of Brahmi script in derivative writing systems.” In: *Written Language & Literacy* 15.1, pp. 1–25.
- (2020). *Линейный и эмблематический принципы в письме и в языке* [*Linear and Emblematic Principles in Writing and in Language*]. Moscow: RSUH.
- (2021). “On the Typology of Writing Systems.” In: *Proceedings of Grapholinguistics in the 21st Century 2020*. Ed. by Y. Haralambous. Brest: Fluxus Editions, pp. 805–824.
- (2023 [2015]). *История и теория письма* [*History and theory of Writing*]. 4th ed. <https://www.litres.ru/book/ludmila-fedorova/istoriya-i-teoriya-pisma-11645652/>. Moscow: Flinta-Nauka.
- Galarza, Joaquín and Rubén Maldonado Rojas (1986). *Amatl, amoxtlí. El papel, el libro. Los códices mesoamericanos*. México: SEIT-ENAH.
- Gelb, Ignace J. (1963). *A Study of Writing*. Chicago-London: The University of Chicago Press.
- Harris, Roy (1986). *The Origin of Writing*. Italian. new augmented Italian edition 1998. London: Duckworth.
- (1990). “The Semiology of Textualization.” In: *The Foundation of Linguistic Theory. Selected writings of Roy Harris*. Ed. by N. Love. London-New York: Routledge, pp. 210–226.
- (2000). *Rethinking Writing*. London: The Athlone Press.
- Hjelmslev, Louis (1973). “The Basic Structure of Language.” In: *Travaux du Cercle Linguistique de Copenhague XIV*. original text dated to 1947, pp. 119–156.
- Klinkenberg, Jean-Marie and Stéphane Polis (2018). “On Scripturology.” In: *Signata* 9, pp. 57–102.
- Meletis, Dimitrios (2020). *The Nature of Writing. A Theory of Grapholinguistics*. Vol. 3. Grapholinguistics and Its Applications. Brest: Fluxus Editions.
- Mounin, Georges (1970). “Le blason.” In: *Introduction à la sémiologie*. Paris: Éditions de Minuit, pp. 109–115.
- Peñafiel, Antonio (1885). *Nombres geográfico de México*. México: Oficina Tipográfica de la Secretaría del Fomento.
- (1895). *Nomenclatura geográfica de México. Primera parte*. México: Oficina Tipográfica de la Secretaría del Fomento.
- Perondi, Luciano (2012). *Sinsemie. Scrittura nello spazio*. Viterbo: Stampa Alternativa & Graffiti.
- Perondi, Luciano and Antonio Perri (2018). “Framing space in Aztec writing. The *Codex Mendoza* as a model of transposition and beyond.” In: *XYdigitale* 5, pp. 40–53.

- Perri, Antonio (1994). *Il Codex Mendoza e le due paleografie*. Bologna: Clueb.
- (2001). “Scrittura, immagine e linearità. Il caso azteco.” In: *Notizie ALAP* 11, pp. 10–13.
- (2007). “Al di là della tecnologia, la scrittura. Il caso Unicode.” In: *Annali. Università Suor Orsola Benincasa* 2, pp. 725–748.
- Vaillant, Pascal (1999). *Sémiotique des langages d’icônes*. Paris: Honoré Champion.
- Vygotsky Lev, S. (1982). “Thinking and speech.” In: *Collected works*. Vol. 2. New York: Plenum Press.
- Whittaker, Gordon (2021). *Deciphering Aztec Hieroglyphs*. Oakland, CA: University of California Press.
- Wimmer, Alexis (2006). *Dictionnaire de la langue nahuatl classique*. [On the base of: Siméon, Rémi. *Le Dictionnaire de la langue nahuatl ou mexicaine*. 1885]. <http://sites.estvideo.net/malinal/nahuatl.page.html>.



# Perceptual Disfluency Through Hard-to-Read Fonts. Is There a Satisfactory Explanation?

Mary C. Dyson

*Abstract.* Research on perceptual disfluency has demonstrated an apparent memory advantage for hard-to-read (less legible) text. This paper explores the evidence, outlines alternative theories, and discusses the locus of the effect. In particular, accounts which propose a metacognitive explanation are contrasted with those which focus on earlier levels in the reading process: letter and word recognition. The reviewed studies illustrate the unreliability of perceptual disfluency effects and confirm the need for further exploration of boundary conditions and moderating factors.


## Introduction

Fluency or disfluency is variously described as a subjective experience of ease or difficulty associated with cognitive tasks (e.g., Diemand-Yauman, Oppenheimer, and Vaughan 2011) or mental processes (e.g., Oppenheimer 2008). When applied to reading, words may be made harder to read through, for example, the use of complicated language (lexical disfluency) or a less legible font or handwriting (perceptual disfluency).

This paper focuses on perceptual disfluency (sometimes described as simply disfluency) as this concerns the graphic representation of language. Studies of perceptual disfluency include manipulations of reading material that change the typeface or variant (e.g., from roman to italic), vary the contrast (e.g., from black to grey type), and compare handwriting to type. All these studies use the Latin script.<sup>1</sup>

The article by Diemand-Yauman, Oppenheimer, and Vaughan (2011), published in the journal *Cognition*, attracted a lot of media attention, as it

---

Mary C. Dyson  0000-0002-0920-4312  
Department of Typography & Graphic Communication  
University of Reading, Reading, RG6 6BZ, UK  
E-mail: m.dyson@reading.ac.uk

1. I am aware of only one study that used material in Hebrew (Sidi, Ophir, and Ackerman, 2016) in which participants were required to solve misleading maths problems.

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 101–109. <https://doi.org/10.36824/2022-graf-dyso>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

presented empirical evidence for better recall of hard-to-read materials compared with easy-to-read materials. These results were found in a classroom environment as well as a laboratory setting, which perhaps contributed to their impact.

As a psychologist working in the field of typography, I find the apparent memory advantage of material that is hard to read difficult to reconcile with a body of legibility research which promotes ease of reading. Having made my bias explicit, this paper explores the evidence for disfluency effects, alternative theories, and the locus of the effect.

## Replications

Since the publication of the article by Diemand-Yauman, Oppenheimer, and Vaughan (2011) reporting these counter-intuitive results, various replications have been attempted and boundary conditions or moderating factors explored (e.g., Kühl and Eitel 2016). These studies were in response to the paucity of studies confirming the basic effect. Based on a meta-analysis of twenty-five empirical studies, the generality of the disfluency effect with respect to learning has been questioned (e.g., Xie, Zhou, and Liu 2018).

Similarly, the creation of a new font, Sans Forgetica,<sup>2</sup> designed to be harder to read to boost memory, has been put to empirical test by various researchers (Geller, Davis, and Peterson, 2020; Taylor et al., 2020; Dyson and Březina, 2021; Eskenazi and Nix, 2021; Wetzler, Pyke, and Werner, 2021). The conclusions are consistent in failing to demonstrate an advantage: ‘Although Sans Forgetica is novel and hard to read, its effects might well end there’ (Taylor et al., 2020, p. 6); Sans Forgetica is not desirable for learning (Geller, Davis, and Peterson, 2020); disfluent fonts are not always desirable difficulties (Wetzler, Pyke, and Werner, 2021).

Given the inconsistent findings,<sup>3</sup> the theoretical underpinnings of perceptual disfluency could benefit from closer examination.

## Metacognitive theory

Diemand-Yauman, Oppenheimer, and Vaughan (2011) refer to the original metacognitive theory of disfluency (e.g., Alter, Oppenheimer, Epley, and Eyre 2007), which is also used to frame the studies published in a special issue of *Metacognition and Learning*, edited by Kühl and Eitel

---

2. <https://web.archive.org/web/20200611220322/http://sansforgetica.rmit/>

3. Some of these are summarised in Dyson (2020).

(2016). When applied to perceptual disfluency and memory, this explanation posits that a reader recognises a word, then perceives the difficulty (a metacognitive cue), puts more effort into processing the word, and therefore is more likely to remember what they have read. The difficulty in recognising the letters (in a hard-to-read font) and identifying a word is a perceptual difficulty, yet this perceptual process is explained in cognitive or metacognitive terms.

This theory of disfluency has been linked with two different psychological accounts of processing:

- Typically, disfluency references James (1950) who claimed that we have two processing systems: one is quick, effortless, and intuitive; another is slow, effortful, analytic, and deliberate. If the content of what we read is simple, but in a hard-to-read font, we may be tricked into using the second system which processes more deeply.
- Geller (2017, p. 11) relates the metacognitive theory to the level of processing framework proposed by Craik and Lockhart (1972) whereby words that are processed to deeper levels (i.e., semantic) are better remembered.

## Alternatives to metacognitive theory

More recently, studies have proposed and tested alternative accounts of perceptual fluency, perhaps prompted by the difficulties in replicating the findings of better performance with disfluent material.<sup>4</sup>

The locus of the disfluency effect has been explicitly questioned by Geller (2017). Drawing on the Interactive-Activation model of visual word recognition (McClelland and Rumelhart, 1981), Geller characterises the level of theoretical mechanisms as pre-lexical, lexical, or post-lexical. When reading disfluent text, the nature of additional activity required at each level is described:

- At the pre-lexical level, where parallel letter recognition occurs,<sup>5</sup> hard-to-read text would require additional processing to identify the letters.
- At the lexical level, more feedback is needed from the word level down to the letter level to identify the letters.
- At the post-lexical level, more feedback is needed from the semantic level down to the word level—the metacognitive theory.

---

4. Although fluent or disfluent relates to the processing of the material, rather than the material itself, researchers often use the term to describe the material. This also applies to the use of the term 'legible text', referring to ease of reading.

5. There is broad agreement amongst reading researchers that word recognition is based on parallel letter recognition (Larson, 2005).

In his thesis, Geller (2017) explores the theories associated with each level of processing and examines the evidence for each of these.

### Pre-lexical: encoding effort hypothesis

The encoding effort hypothesis proposes that the effort required to identify items enhances memory for these items. One of the experiments conducted by Hirshman, Trembath, and Mulligan (1994) varies the contrast between text and background with either grey letters on a black background or white letters on a black background. Although identification of words in grey was more difficult (took longer), recall was comparable in both conditions.

### Lexical: compensatory processing account

Geller, Still, Dark, and Carpenter (2018) introduce the compensatory processing account as a possible explanation for disfluency results. This account is used by Hirshman, Trembath, and Mulligan (1994) to explain their finding that visual masking of words enhances memory. They conclude that higher level processing is compensating for visual processing difficulties and the additional activity is improving memory.

A similar emphasis on word-level processing is proposed by Wetzel, Pyke, and Werner (2021), but in this case, to explain the lack of a memory benefit from the disfluent font (*Sans Forgetica*). They propose that a disfluent font increases the demands on orthographic processing but does not help, and may even impair, semantic relational processing by slowing down reading. Being aware of the perceptual difficulty (metacognition) did not improve recall.

Handwriting also provides a means of exploring the use of top-down processes as there is an inherent physical variability in letter forms that is not found in a fluent font. A study comparing handwriting to Courier New font found that various lexical effects (word frequency, consistency, and imageability) were enhanced with handwriting compared with Courier New (Barnhart and Goldinger, 2010). They propose that handwriting requires greater use of top-down processing because it departs from the 'more prototypical word forms' (p. 921). The notion of a prototype fits with typographers' belief that typeface familiarity is important to legibility. This prototype hypothesis has been investigated by comparing fonts with common letter shapes and uncommon letter shapes (Beier and Larson, 2013).<sup>6</sup>

---

6. They found no difference in speed of reading between common and uncommon letter shapes, but participants disliked the uncommon shapes.



An alternative explanation for handwriting needing more top-down processing is that the letters are noisy, ambiguous forms, rather than departing from a prototype. These two hypotheses were tested by Perea, Gil-López, Beléndez, and Carreiras (2016) by comparing difficult-to-read and easy-to-read handwriting with the typeface Century. They found that handwriting was read more slowly, and less accurately than Century. However, there was no difference in lexical effects (word frequency) between the easy-to-read handwriting and Century, whereas harder to read handwriting did show a word frequency effect. The quality of the handwritten words is therefore important in moderating the use of top-down processes.

### Load theories

Another way of describing the different levels is in terms of load theory where some researchers have distinguished between sensory, perceptual, and cognitive load in the context of disfluency (Marsh et al., 2018; Hao and Conway, 2022).

Cognitive load has been proposed as an alternative to disfluency theory (Kühl and Eitel, 2016). According to cognitive load theory, learning material should be designed to decrease demands on working memory which has limited capacity. This theory therefore proposes the use of legible or fluent texts to support ease of reading. Their series of four studies produced contradictory results, failing to confirm either cognitive load or disfluency theory. This led them to conclude that the less legible text layout may have increased the perceptual load, rather than cognitive load.

A study that considers the potential effects of different types of load looked at the disruptive effect of background speech on reading comprehension (Hao and Conway, 2022). They found that a disfluent font improved comprehension but there was no benefit from the disfluent font with background speech. The authors argue that a disfluent font introduces a perceptual load. Citing Lavie and De Fockert (2003), they query the extent to which texts with reduced contrast, or smaller font size, can be described as perceptually disfluent as these manipulations may introduce a sensory load, but not a perceptual load. They distinguish between these sensory degradations and a hard-to-read font which may increase perceptual load because additional perceptual operations are required.

Also looking at attention and task engagement, Faber, Mills, Kopp, and D'Mello (2017) investigated the effect of a (supposedly) disfluent font (Comic Sans, italic, grey) on mind wandering and comprehension when reading a text about scientific research. They found less mind wandering with Sans Forgetica but no effect on comprehension and sug-

gest that disfluency may impose an extraneous cognitive load, offsetting the advantage of less mind wandering.

There seems to be disagreement on whether disfluent fonts introduce an additional perceptual or cognitive load. An insight into which stage of the reading process may be affected by background speech comes from eye movement recordings (Vasilev et al., 2019). They found that background intelligible speech only affects the post-lexical stage of processing when readers integrate words into sentences. With the proviso that Vasilev et al. did not include a disfluency manipulation, this finding may contribute to explaining why Hao and Conway (2022) found no shielding effect from the disfluent font in background speech. They claim that the disfluent font introduces a perceptual load, and a high perceptual load filters irrelevant information as the perceptual processes are fully engaged by task-relevant information. If the background speech distraction is indeed affecting a later stage of processing, there will be no shield against the distraction from perceptual disfluency.

## Discussion

Unfortunately, a satisfactory explanation for perceptual disfluency has not emerged from the empirical research described above, and further questions are raised. On the one hand, various accounts seek to explain how additional processing enhances memory, and on the other hand theories of extraneous load predict impaired performance. Both strands incorporate different levels of the reading process: pre-lexical, lexical, post-lexical and sensory, perceptual, and cognitive. There is some convergence of evidence that disfluent text requires extra processing at the word level but uncertainty as to whether this aids or impedes memory. This may depend on the reader as a disfluent font may not improve performance unless they have sufficient working memory capacity (Lehmann, Goussios, and Seufert, 2016).

Of particular importance from a verbal graphic language perspective is the need to establish empirically, rather than assume, whether a font used in a study is hard-to-read. The discrepant results, including different qualities of handwriting (Perea, Gil-López, Beléndez, and Carreiras, 2016), highlight the importance of attempting to calibrate degrees of disfluency. There is some evidence for a reverse U-shape curve when plotting performance against level of disfluency (Seufert, Wagner, and Westphal, 2017): learning is improved up to a certain level of disfluency but increasing beyond this point impairs learning. We currently have no means of mapping different fonts or font variants (bold, italic) on a fluency or legibility scale to search for an optimum level of disfluency. But, at the very least, all studies could include participant's comparative

judgements of legibility of test material to validate perceived differences between material labelled as hard- or easy-to read.

The moderation of the use of top-down processes by the quality of handwriting may shed some light on the failure of Sans Forgetica, and other fonts, to display disfluency effects. Perea, Gil-López, Beléndez, and Carreiras (2016) describe the normalisation process that occurs with easy-to-read handwriting, where we tune into the idiosyncrasies of the handwriting. There is a similar process with fonts, described as ‘font tuning’ where consistency increases letter identification efficiency (Sanocki and Dyson, 2012). With Sans Forgetica, it may be possible to tune into the unusual letter forms, given some exposure. Beier and Larson (2013) confirmed that twenty minutes reading a font with uncommon letter shapes increased speed of reading.

In conclusion, it is reassuring that the early stages of reading (from letter to word) are no longer ignored in explanations of perceptual disfluency. Although the dispute between the beneficial effects of disfluency versus legibility is not yet resolved, useful questions have been asked.

## References

- Alter, A. L. et al. (2007). “Overcoming intuition: Metacognitive difficulty activates analytic reasoning.” In: *Journal of Experimental Psychology: General* 136, pp. 569–576.
- Barnhart, Anthony S. and Stephen D. Goldinger (2010). “Interpreting chicken-scratch: lexical access for handwritten words.” In: *Journal of Experimental Psychology: Human Perception & Performance* 36.4, pp. 906–923.
- Beier, Sofie and Kevin Larson (2013). “How does typeface familiarity affect reading performance and reader preference?” In: *Information Design Journal* 20.1, pp. 16–31.
- Craik, F. I. M. and R. S. Lockhart (1972). “Levels of processing—framework for memory research.” In: *Journal of Verbal Learning and Verbal Behavior* 11.6, pp. 671–684.
- Diemand-Yauman, Connor, Daniel M. Oppenheimer, and Erikka B. Vaughan (2011). “Fortune favors the bold (and the Italicized): effects of disfluency on educational outcomes.” In: *Cognition* 118.1, pp. 111–5.
- Dyson, M.C. (2020). “Does perceptual disfluency theory represent a significant challenge to a legibility researcher?” In: *Hyphen* 12.18, pp. 17–35.
- Dyson, Mary C and David Březina (2021). “Exploring disfluency: Are designers too sensitive to harder-to-read typefaces?” In: *Design Regression*. <https://designregression.com/research/exploring-disfluency-are-designers-too-sensitive-to-harder-to-read-typefaces>.

- Eitel, A. et al. (2014). "Disfluency meets cognitive load in multimedia learning: Does harder-to-read mean better-to-understand?" In: *Applied Cognitive Psychology* 28.4, pp. 488–501.
- Eskenazi, Michael A. and Bailey Nix (2021). "Individual differences in the desirable difficulty effect during lexical acquisition." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 47.1, pp. 45–52.
- Faber, Myrthe et al. (2017). "The effect of disfluency on mind wandering during text comprehension." In: *Psychonomic Bulletin Review* 24.3, pp. 914–919.
- Geller, Jason (2017). "Would disfluency by any other name still be disfluent? Examining the boundary conditions of the disfluency effect." PhD thesis. Iowa State.
- Geller, Jason, Sara D. Davis, and Daniel J. Peterson (2020). "Sans Forgetica is not desirable for learning." In: *Memory* 28.8, pp. 957–967.
- Geller, Jason et al. (2018). "Would disfluency by any other name still be disfluent? Examining the disfluency effect with cursive handwriting." In: *Memory & Cognition* 46.7, pp. 1109–1126.
- Hao, Han and Andrew R. A. Conway (2022). "The impact of auditory distraction on reading comprehension: An individual differences investigation." In: *Memory & Cognition* 50.4, pp. 852–863.
- Hirshman, Elliot, Dawn Trembath, and Neil Mulligan (1994). "Theoretical implications of the mnemonic benefits of perceptual interference." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20.3, pp. 608–620.
- James, William (1950). *The principles of psychology*. Original work published 1890. New York: Dover.
- Kühl, Tim and Alexander Eitel (2016). "Effects of disfluency on cognitive and metacognitive processes and outcomes." In: *Metacognition and Learning* 11.1, pp. 1–13.
- Larson, Kevin (2005). "The science of word recognition or how I learned to stop worrying and love the bouma." In: *Typo* 13, pp. 2–11.
- Lavie, N. and J. W. De Fockert (2003). "Contrasting effects of sensory limits and capacity limits in visual selective attention." In: *Perception & Psychophysics* 65.2, pp. 202–212.
- Lehmann, J., C. Goussios, and T. Seufert (2016). "Working memory capacity and disfluency effect: an aptitude-treatment-interaction study." In: *Metacognition and Learning* 11.1, pp. 89–105.
- Marsh, John E. et al. (2018). "Why are background telephone conversations distracting?" In: *Journal of Experimental Psychology: Applied* 24.2, pp. 222–235.
- McClelland, J. L. and D. E. Rumelhart (1981). "An interactive activation model of context effects in letter perception. 1. An account of basic findings." In: *Psychological Review* 88.5, pp. 375–407.
- Oppenheimer, D. M. (2008). "The secret life of fluency." In: *Trends in Cognitive Sciences* 12.6, pp. 237–241.

- Perea, Manuel et al. (2016). "Do handwritten words magnify lexical effects in visual word recognition?" In: *The Quarterly Journal of Experimental Psychology* 69.8, pp. 1631–1647.
- Sanocki, Thomas and Mary C. Dyson (2012). "Letter processing and font information during reading: Beyond distinctiveness, where vision meets design." In: *Attention Perception Psychophysics* 74.1, pp. 132–145.
- Seufert, Tina, Felix Wagner, and Julia Westphal (2017). "The effects of different levels of disfluency on learning outcomes and cognitive load." In: *Instructional Science* 45.2, pp. 221–238.
- Sidi, Yael, Yael Ophir, and Rakefet Ackerman (2016). "Generalizing screen inferiority—does the medium, screen versus paper, affect performance even with brief tasks?" In: *Metacognition and Learning* 11.1, pp. 15–33.
- Taylor, Andrea et al. (2020). "Disfluent difficulties are not desirable difficulties: the (lack of) effect of Sans Forgetica on memory." In: *Memory* 28.7, pp. 850–857.
- Vasilev, Martin R. et al. (2019). "Reading is disrupted by intelligible background speech: evidence from eye-tracking." In: *Journal of Experimental Psychology-Human Perception and Performance* 45.11, pp. 1484–1512.
- Wetzler, Elizabeth L., Aryn A. Pyke, and Adam Werner (2021). "Sans Forgetica is not the "font" of knowledge: Disfluent fonts are not always desirable difficulties." In: *Sage Open* 11.4.
- Xie, Heping, Zongkui Zhou, and Qingqi Liu (2018). "Null effects of perceptual disfluency on learning outcomes in a text-based educational context: A meta-analysis." In: *Educational Psychology Review* 30.3, pp. 745–771.



# Asemic Writing. Homebound

Christine Kettaneh


*Abstract.* To free ourselves from the confinement of home, the artist proposes a journey of forgetting the limiting meaning of home. The journey starts with testing the elasticities of letterforms and signs, breaking them beyond legibility, while exploring spectrums between word and image, until it reaches the territories of asemiosis. Once freed from meaning, the search continues along the threads of asemic writing triggering questions and affects. A step deeper along the meaningless but remarkable traces takes the artist into nature, where she realizes that the escape from home has taken her back home, the original home.

## 1. Forgetting

During the Covid-19 pandemic, confinement made many of us question the notion of home and the limits of our space. Home is not just the physical space we inhabit. It is the domain of our intimate being. It is where we dream and make memories. So when that space is questioned or attacked, we become anxious. We fidget and become restless pacing the space back and forth, opening and closing windows and doors... until we realize that perhaps the only way out of that unrest is through forgetting language, the system or the beliefs that define and limit our notion of home. That way we are open of resetting and finding new meanings of home. We may then be open to dreaming differently and remembering differently. So maybe we need to forget the word Home, starting with the letter H.

With that intention in mind, I developed an artwork in the form of an animation called “Limits of H” (Fig. 1). In devising it, I ran visual digital tests on the letter H with code. I wanted to see how far I needed to change the basic segments of that letter until I no longer recognized it as H. As the animation runs, there is a growing understanding of a

---

Christine Kettaneh  0000-0003-4099-9990  
Office GB 820, Floor 8, Gezairi Buiding, School of Architecture and Design Lebanese  
American University, Beirut Campus, Lebanon  
E-mail: christinekettaneh@gmail.com

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 111–134. <https://doi.org/10.36824/2022-graf-kett>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

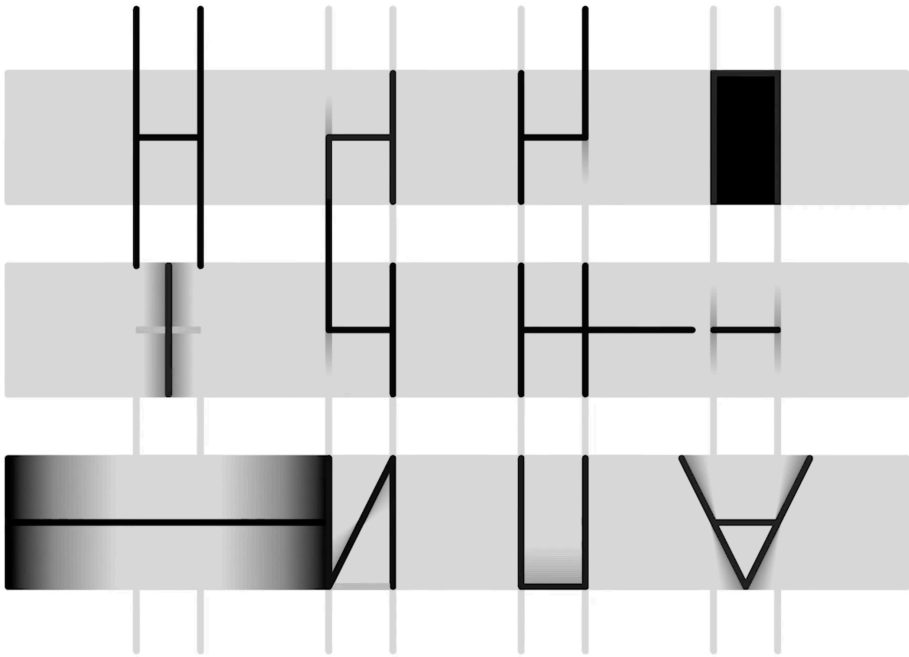


FIGURE 1. Still from “Limits of H,” Christine Kettaneh, 2021, video, 9 mins

repetitive going from what we recognize as independent Hs, to something more visually abstract yet somehow more interconnected. However, one cannot help capturing random instances of alternative recognizable symbols like A, V or even ancient Phoenician letters (Kettaneh, 2021).

I wonder how many of Changizi’s (2006) configuration types the letter H is touching upon as it transitions from its original verbal to its different more abstract final forms. In his study, Changizi identifies 36 different configuration types across 100 writing systems over human history, Chinese characters, and nonlinguistic symbols, (while confining the samples to characters of three or fewer strokes) (Fig. 9). Each configuration type captures a strong distinct topological identity that is invariant to various geometrical shape variations like variations in relative orientations, lengths, and shapes of the segments or the orientation of the overall character. As my set of Hs animate, I realize that some intermediate forms linger in one configuration as others jump into another one while still others jump outside of that catalogue of configurations all together. So I suppose, the animations of ‘Limits of H’ are most probably flickering in and out of that catalogue as our minds attempt to read alternating instances of legibility and illegibility. I then did a



similar test on the letter ب in Arabic developing “Limits of ب,” where ب is the first letter of Beit (‘Home’ in Arabic) (Fig. 10).

## 2. Kineticism

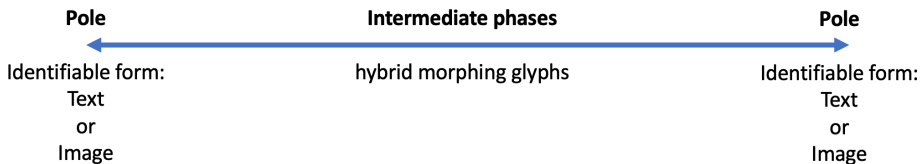


FIGURE 2. Diagram inspired from Barbara Brownie’s continuum of fluidity

Reflecting on the limits of H and ب further and focusing on the elasticity of the letters that is expressed, led me to Barbara Brownie’s recent studies on kinetic typography (2015). She says that most research in temporal media talk about motion or displacement of whole letters or words along the screen; but they overlook the instances where temporal media has allowed for exploring the malleability of the individual letter. The type designer’s role is to create and transform letterforms. In print, such workout remains hidden and what we see or use is the end result. But temporal media has allowed for this work-out to be visible: we can see letters on the screen being created, sculpted and transformed. Brownie calls those changes local kineticism.

However, what Brownie is mostly interested in is ‘fluidity,’ the extremes of local kineticism, where deformations affect the identity of the letterform allowing for a transformation of nature and meaning. The verbal identity might transform into another verbal identity or into an entirely new pictorial or abstract identity. She calls the identities poles. Usually at every pole, we have a form that can be easily recognized as text or image. During the transformation one identity is lost and strange hybrid nonsemantic forms arise before the next pole is reached. However, the meaning of the artifact is not complete at any one pole and is not the sum of the two poles. The meaning is more complex and can only unravel gradually across time from text to image, text to text or pole to pole. My visualization of this fluidity can be seen in Fig. 2. For Brownie’s examples, check Fig. 11 that draws a man transforming into the letter “x” and the letter “k” transforming into the letter “m.”

So the asemiosis during the transformation is significant because it resembles two things:

1. A learning experience: The unidentifiable glyphs in the intermediate stages create discomfort because the spectator experiences a phase

of unsettling illiteracy. So those glyphs provoke the spectator's anticipation for the emergence of more familiar signs. So in a way the intermediate phases of transformation prepare the viewer to become a reader and vice versa. When the meaning emerges at the next pole there is a moment of relief and satisfaction of newly acquired knowledge. So fluid typography is like learning: "As asemic signs become legible, new knowledge and understanding is granted to the reader, as if he or she has just learned to read." (Brownie, 2015, p. 57)

A great example showing how asemiosis can support the learning experience is Colleen Ellis's work in ABCing (Ellis, 2010). She breaks down the alphabet by breaking down not the letter itself but the space around it. She then moves and rearranges the pieces so they form a new sign reflecting the meaning of a word that starts with the original letter. It reminds us of the experience of a child learning the alphabet, yet now taken at a second level: our adult eyes already trained to see the alphabet, Colleen guides us to find meaning outside it. The learning is facilitated through her animations that accompany the book. The animations show the process of unlearning to relearning as the meaning disappears with the letter and reemerges in a new form.

In Fig. 12, O breaks down into an organic shape: "a shape relating to, or suggestive of, the natural world or living organisms. [<Latin *organicus* <Greek *ὀργανικός*, "of or pertaining to an organ" + < Old English *gesceap*, "creation, form, destiny".]

2. A live experience: Most of our human experiences are analogue "involving graded relationships on a continuum." So when we try to express it in words we fall short because words do not operate in a continuum. By naming things "we reduce the continuous to the discrete" and we end up perceiving our experiences as binary. On the other hand, fluid artifacts and their transformations give as much importance to the poles as the variations happening in between them allowing for a continuous experience.

Fig. 13 shows a good example: one of Dan Waber's strings called "Argument" (2005). It presents a single string which repeatedly reforms itself between two words: yes and no. The clear yes becomes uncertain before it becomes a clear no and vice versa. Yes and no are binary opposites but "Argument" bounds them across time with the string and hence "presents them on an analogue continuum." (Brownie, 2015, pp. 87–88)

### 3. Asemiosis

The intermediate unidentifiable glyphs in the kinetic works arise as a consequence of fluidity. However, signs which function in similar ways appear in static media, and have been named by Tim Gaze and Jim Left-

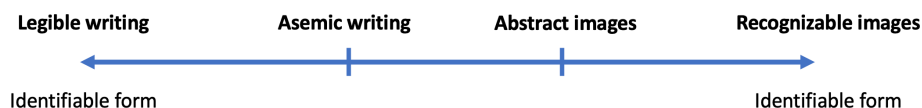


FIGURE 3. Diagram inspired from Tim Gaze's spectrum between text and image

wich, two visual poets, as 'asemic writing.' They coined this name in the 1990s when there was a surge of asemic works being shared over the internet. But asemic writing actually started long before it was named as such like in the poetry of Henri Michaux, writings of Roland Barthes and paintings of Cy Twombly.

At its simplest, asemic writing is, according to Tim Gaze, anything which looks like writing through their shape or organization, but in which the person viewing can't read any words. It is a kind of "writing without Language." Leftwich explained in a letter to Gaze written on January 27, 1998: "A seme is a unit of meaning, or the smallest unit of meaning (also known as a sememe, analogous with phoneme). An asemic text, then, might be involved with units of language for reasons other than that of producing meaning." (Schwenger, 2019, p. 1)

In his reflections on asemiosis,<sup>1</sup> Gaze (2011) suggests a continuum that exists between image and legible writing or between image and text. At one end of the continuum lies legible writing; at the other end lies recognizable images; and in between the two ends, closer to the legible writing, lies asemic writing and then abstract images. Gaze's spectrum might look as depicted in Fig. 3 and might resemble the continuum that fluid artifacts perform. However in fluid artifacts, asemiosis is temporary and hence the uncomfortable phase of illegibility is temporary. But in asemic writing in static media, the illegibility is given fixed and does not promise a solution. So if asemic writing leaves us frustrated, then why does it still appeal to us and why do we still produce it?

Interviewed by Asymptote, Michael Jacobson says that "asemic writing offers meaning by way of aesthetic intuition, and not by verbal expression." Even if it is illegible, it is still attractive to the eye because it has an open semantic form that can relate to all words, colors and music. More importantly, it can relate to emotions that cannot be expressed with words. So asemic writing fills in a need and is also international in its mission. It is active beyond the language of the author or reader.

Michael Jacobson is known for his asemic works and his online gallery, The New Post-Literate, a weblog that "explores asemic writing in relation to post-literate culture." There is an interesting letter in his

1. *Note by the Editor:* Brownie (2014b) uses the term "asemisis" in the title and abstract of her paper. This term is not a neoclassical compound since the form \*semisis does not exist. We will use the term "asemiosis" (privative ἀ- and σημείωσις) instead to designate the same concept.

weblog. It is from Cecil Touchon, also one of the main contemporary names in asemic writing, addressing Peter Schwenger. In his letter, he agrees with Schwenger that legible writing disappears physically upon reading it, and is experienced more as a mental dialog. In that way, he adds, words act more like a 'delivery system.' So one of the reasons why Touchon does asemic works "is to present the actual writing itself as its own concrete, unique reality rather than being representative of something else." He explains that this argument is the same "that stems from abstract or concrete art." (The New Post-literate n.d.)

Touchon then reflects on the experience of reading asemic works. He says that words are signposts that direct the eyes to read the text sequentially word by word, line by line, in order to understand the idea or narrative. So when writing loses its words, the eyes are left wandering around at their own whim. The focus is lost and that might look like disengagement. However, the reality is perhaps that the reading is just different. Devoid of words, we are now looking for patterns, energies and textures enjoying the work as a whole, discovering new things or layers at every reading. This reading experience resembles the reading of an abstract artwork or the appreciation of a musical piece allowing us the liberty to flow in and out of focus with every reading, while encouraging multiple readings.

According to Gaze, reading—and not writing—determines whether a piece of writing is asemic or not. Gaze implies that asemiosis is subjective; if a reader is not able to read a text, then the text is considered at that point in time, as asemic. But it might not be asemic to somebody else who is able to read it. Asemiosis proves also to be subjective along his suggested continuum between image and text.

One person sees a picture of a house (recognizable image); another sees a bundle of lines (abstract image). One person can read a piece of graffiti (legible writing); another can't (asemic writing). One person sees an unknown species of writing (asemic writing); another sees spaghetti (abstract image). (Gaze, 2011)

Schwenger (2019) discusses another aspect involved in the reading experience of asemic writing. When faced with an asemic piece, we might notice our first impression: an expectation that the text is legible. Upon our failure to read it, some of us might take the piece lightly and impatiently disengage. Others might resist its illegibility and try relentlessly to decode it or translate it desiring that the text rewards them with meaning. Either way, both reactions may reflect to us our addiction to verbalizing and our dependency on logical orders.

Although there is a lot of asemic work being produced and circulated, and an increasing interest in it, yet there is not much written about it. One of my main references was Peter Schwenger's book: *Asemic. The Art of Writing* (2019) which can be considered as the first map

and critical study of this fascinating field. Schwenger discusses the works of three asemic ancestors: Henri Michaux, Roland Barthes and Cy Twombly. Understanding their approaches lays ground to most asemic works.

### 3.1. Henri Michaux, *Mouvement*

To make visible the interior sentence, the sentence without words, the cord that unrolls itself infinitely, sinuously, and deep within accompanies everything that presents itself, outside as well as inside. (Henri Michaux)

Henri Michaux is known for both his literary and art works. He aims in both fields to push beyond conventions towards what he calls “the space within, or beyond.” He feels that words are limiting because they are kind of images but restrictive ones. So he desires to build “sentences without words,” sentences that escape translation. This leads him to an asemic practice that focuses on movement. He wants continuity and change devoid of the stop signs of words. Not surprisingly, he is very much influenced by dance and the language of the body.

Fig. 14 is a work by Henri Michaux. It is from his book *Mouvements* (1951/1982) which is a book of markings that was created by improvised impulses: movements of the hand and accidents of ink. He calls his asemic forms not as shapes but as interior gestures. These gestures do not convey thoughts or stable signs but rather reflect an interior tempo. This interior tempo is emotion, which is part of our response to sensation or perception. Our emotions accompany our first vague forms of our ideas. So his gestures reflect not thought but what precedes thought: an expression of our primal desire. (ibid.)

### 3.2. Roland Barthes, *Contre-écriture*

Roland Barthes is influenced by Henri Michaux and several others when he also tries to avoid meaning in order to unlock the power of the asemic. Fig. 7 is a selection from Barthes’s “contre-écriture,” published in 1976 in a journal (Barthes 1976, and Onnen 2008, p. 27). His author’s note reads: “If my graphisms are illegible, it is precisely in order to say No to commentary.” This is not a reflection of insecurity about his work but rather a hint about commentary’s fundamental nature: “For commentary endlessly extends language; it is in the service of an impossible quest to extract the last, the final, drop of meaning.” (Schwenger, 2019, p. 32)

He has an “almost obsessive relation to writing instruments.” For Barthes, writing is a sensual act; he is very much interested about the

muscular act of tracing letters, its physicality, its scription, and the resulting materiality of accidental ink blots, gesture painting or unconscious doodling. He is also interested in the speed of writing and how it conveys the author's style. He suggests that speedy writing can reveal a "kinetic relationship between the head and the hand." "In this relationship the head does not automatically have priority: it may be dictated to by the hand quite as readily as the other way around." (Schwenger, 2019, p. 37)

He chooses to experiment with asemic writing as an anti-mythological action. He wants to overturn the old myth that assumes thought precedes language and that language is only an instrument to transmit those thoughts, ideas or information. He is in line with Saussure's proposition that "without language, thought is a vague, uncharted nebula. There are no pre-existing ideas, and nothing is distinct before the appearance of language." Hence the sign or writing is the condition of thought, not its instrument, medium, or expression. (Badmington, 2008, p. 89)

All the materials and material act of writing that he is interested in: the hand, the pen, the paper are usually overlooked because we are conditioned to prioritize meaning in our reading. So to make writing visible in its truth, Barthes suggests that writing needs to be illegible. It is the only way that the graphic element would reclaim its primacy. He calls his asemic writing graphism in order to bridge the gap between writing and painting which he believes are not fundamentally different.

### 3.3. Cy Twombly, *Letter of Resignation*

The line is the feeling, from a soft thing, a dreamy thing, to something hard, something arid, something lonely, something ending, something beginning. It's like I'm experiencing something frightening, I'm experiencing the thing and I have to be at that state because I'm also going. (Cy Twombly)

We can see those lines of feelings at work in Cy Twombly's "Letter of Resignation," which is a series of thirty-eight drawings, probably done in response to the hostile reviews his works received a year earlier (Fig. 8). Those reviews affected him so much that he takes a break from painting for almost a year. Yet that resignation from art is only a temporary one because just by performing the letter of resignation he is also returning boldly back to art. (Schwenger, 2019)

The emotions he felt must have been of frustration and anger; emotions that were beyond words. So his letters are written with agitation without control, without articulation, without words. Through the violent and agitated markings we can feel the physical venting of the pencil on the page. The writings devoid of verbal meaning return to their primal form as drawing. Only the forward intensity, the leaning, the

cursivity of writing remain. The fact that he has written multiple drafts reinforces the idea that the words come after the feeling. In each attempt, the writer tries to fit words ever more closely to the shape and quality of the feeling.

#### 4. Eco-asemiosiis



FIGURE 4. Still from an animation depicting wormlike movements, from the art film “The Hindwing” (2018) by Christine Kettaneh

“Despite his own artistic ability, Leonardo da Vinci (1452–1519) believed that humans could never create anything “more beautiful, simple, or direct than nature.”” That’s how Robert M. Peck starts his essay “Asemic Writing from the Mouth of a Snail,” in the *Natural History* magazine. In his essay, Peck draws on the artistry in the traveling and eating trails of snails and likens those ethereal patterns to asemic writing. He refers to a photograph (Fig. 15) which he has taken to show the paths created by a common land snail as it feeds on algae off a nutritious surface. He gives a detailed description of the feeding habit that leaves fan-like trails of thin strips that have similarities with the ink drawings of Henri Michaux and Norman Lewis, and the abstract paintings of Cy Twombly. (Peck, 2022)

According to Gaze, “You could say that nature, since time began, has been manifesting asemic writing. It just needs a human to see the writ-

ing, & recognize it". De Villo Sloan named those asemic markings in nature as 'eco-asemics.' (Schwenger, 2019)

I was intrigued myself by trails left by a bark beetle infestation under the barks of my family's historic pine tree (Fig. 16). When a bark beetle overcomes a weak pine tree, I learnt, it makes a nursery inside the bark. The offspring, once hatched, feeds on the soft tissue in the tree making tunnels through the bark. Those tunnels further disrupt the circulation of water and sap of the tree causing its eventual death. I made a short art film called "The Hindwing," (2018) documenting the felling of our family pine tree, while exploring, in parallel, the infestation process. In my studies of the bark beetle, I traced the movement of the body of a larva, the worm stage in the life cycle of the beetle. I took the tracing and used it to animate an abstraction of a larva. I put four of those larvae aligned next to each other unraveling a kind of asemic—or eco-asemic—message through time (Fig. 4).

Some artists hunt for these natural markings and present them directly in their work. Sometimes the artist intervention is minimal like taking a photograph of nature as is but in a specific frame, light or alignment (Fig. 17). Other interventions involve more process, like removing elements from nature and decontextualizing them (Fig. 18–19) or tracing over them (Fig. 20). Sometimes the traces or markings are taken only as studies to influence new works, like in the case of my larvae animation.

In the article "The Structures of Letters and Symbols throughout Human History Are Selected to Match Those Found in Objects in Natural Scenes," Changizi, Zhang, Ye, and Shimojo (2006) demonstrate first that there are empirical regularities governing the topological shapes of human visual signs. He does that by finding strong correlations among the relative frequency of the 36 configurations that were developed across the three classes of visual signs. The results suggest "that the configuration distribution for human visual signs tends to possess a characteristic signature." He then considers an ecological and visual hypothesis for that characteristic signature: that the more common configuration types among visual signs are the more common configuration types among natural scenes. He explains that cultural selection pressure favors configuration types found in natural scenes, because that's "what humans have evolved to be good at visually processing." To test this ecological hypothesis he measures configuration distributions from three classes of natural images: 1. "Ancestral," which consists of photographs of savannas and tribal life. 2. "National Geographic," which consists of photographs of rural and small-town life taken from the National Geographic website. 3. "CGI buildings," which consists of computer-generated realistic images of buildings. The results show that the distributions for the three kinds of environment correlate very highly with one another and more importantly and closely to the signature distri-



bution for human visual signs. The results hence provide evidence to support the ecological hypothesis.

If asemic writing is frustrating because it carries no meaning, eco-asemics is frustrating even more because of the absence of a human author. Without a human author, decoding those signs to satisfy our compulsion for finding a verbal meaning or intention becomes impossible. But just like asemic writing carries the invitation to encounter the physicality of mark making and reading our primal gestures cleared from human language's dictation, eco-asemic writing might be an invitation to consider a language that transcends the human. By offering markings akin to human writing, natural objects might be demanding that we pay attention to them. It might be an invitation to return to nature, the origin upon which we have built our language before we turned it into a human artifice.

## 5. *Tlön, Uqbar, Orbis Tertius*

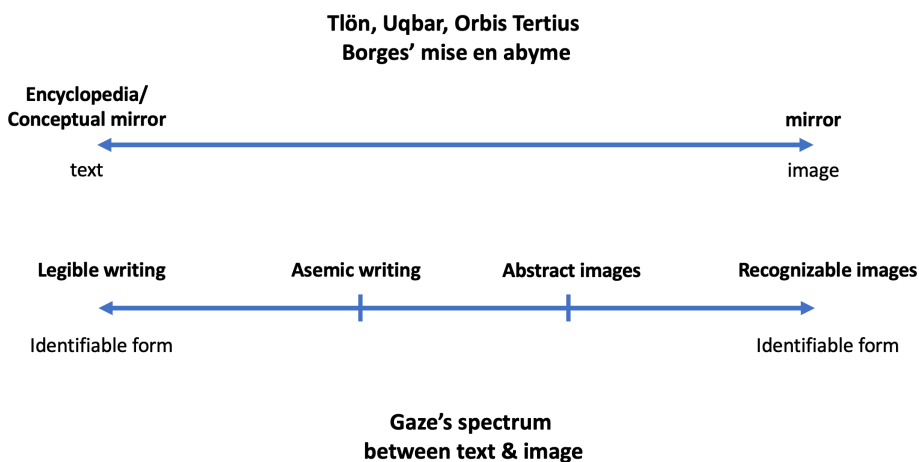


FIGURE 5. Diagram showing the similarity between Borges' mise en abyme and Gaze's continuum between text and image

"Tlön, Uqbar, Orbis Tertius," is the title of a short story written by Jorge Luis Borges in 1940. The story is as difficult as the title sounds. It requires several frustrating readings before you start making a little sense of things.

The story is about an encyclopedia that was written by a secret sect for an imaginary world called Tlön. Tlön is first introduced as a mythical region in a non-existent land called Uqbar. The narrative of the

story is a structure en abîme starting from a reality that matches Encyclopedia Britannica. Reality starts gradually changing as we discover new versions of the encyclopedia, each version having a bigger fictional component. Orbis Tertius is the encyclopedia of Tlön written in the Tlönian language with its own special alphabet. Tlönians are idealists that don't believe in the continuity of objects so their language has no nouns. They have two dialects, one that is based solely on adjectives and one that is based on verbs. Reality is finally threatened to become Tlönian if Orbis Tertius is discovered because that's when the Tlönian language would be adopted and all current languages would be forgotten.

If you read the story carefully you realize that Borges prepared his readers for this mise en abîme when he mentioned at the very start: "I owe the discovery of Uqbar to the conjunction of a mirror and an encyclopedia." So if we think of the encyclopedia as a conceptual verbal mirror of the world, then two mirrors placed in front of each other would lead to a mise en abîme. I find this setup insightful; I think this mise en abîme resembles Tim Gaze's spectrum between image and legible writing. So perhaps asemiosis is a state in a mise en abîme between word and image (Fig. 5). It makes sense then that asemiosis is confusing and frustrating.

Another point from this story which I find significant to the discourse of this paper is the statement that Borges makes towards the end of the story:

How could one do other than submit to Tlön, to the minute and vast evidence of an orderly planet? It is useless to answer that reality is also orderly. Perhaps it is, but in accordance with divine laws—I translate: inhuman laws—which we never quite grasp. Tlön is surely a labyrinth, but it is a labyrinth devised by men, a labyrinth destined to be deciphered by men.

So we have built our languages akin to the original inhuman languages of the world and assigned meaning to them. Humans readily adopted them because they were decipherable. So perhaps Eco-asemic works remind us of the origins of language and asemic writings with their supposed failure to read remind us of our compulsion to assign human meaning.

### 5.1. *Codex Seraphinianus*

"Tlön, Uqbar, Orbis Tertius" inspired many works including "Codex Seraphinianus." It is a 360 pages illustrated encyclopedia of an imaginary world, created by Italian artist, architect and industrial designer Luigi Serafini between 1976 and 1978. The codex is made up of hand-drawn surreal bizarre illustrations divided into two sections. The first section is characterized by the natural world of flora, fauna, anatomies,

and physics. The second is characterized by the various aspects of human life like fashion, architecture, history and foods. The codex is also known for its false writing system. Serafini stated that the writing was asemic and that there was no meaning behind it; he said his experience in writing it was like automatic writing (Fig. 21). “What he wanted his alphabet to convey was the sensation children feel with books they cannot yet understand, although they see that the writing makes sense for adults.” (Babkina, 2015)

Even after such statement, some people still believed the codex could be deciphered and the book’s page-numbering system was decoded by Allan C. Wechsler and Bulgarian linguist Ivan Derzhanski.

## 5.2. The Voynich Manuscript

Maybe the most debated and studied codex of all times—which may have inspired the writing of both “*Tlön, Uqbar, Orbis Tertius*” and “*Codex Seraphinianus*”—remains the enigmatic medieval script, the Voynich Manuscript, that has been carbon-dated to the early 1400s. The Voynich has an interesting history, having been passed through the hands of many scientists, emperors, and collectors. Though the author still remains unknown, studies of its illustrations have hinted that its original purpose is probably medical, including sections akin to “medieval herbals, astrology guides, and bathing manuals.” However, the illustrations look crude and amateurish unlike the more professionally and faithfully drawn plants of the time. More importantly, the illustrations depict botanical impossibilities and surreal imagery which way surpass the little quirks of the medieval herbals. (Hochelaga, 2022)

Adding to the manuscript’s mystery, its 240 pages have been written by hand in an unknown language, referred to as ‘Voynichese’ (Fig. 22). It looks like a European language, reading from left to right, having a 22 letter alphabet combining together to form words. Some tests have shown that the word distribution demonstrates a logic; the spelling reveals some predictable patterns; and some cluster of unique words might hint at keywords belonging to the theme of plants. The presence of an order suggests that the Voynichese behaves like a language; however it is not behaving like any language we know of. Many theories have been developed about the Voynichese. One theory suggests that it is a cipher, a known language in disguise. It has been studied by many cryptographers including codebreakers from both World War I and World War II, but the original language has not been definitely deciphered yet. Another theory suggests that it is a natural language, perhaps a European language that has long been forgotten. But unless we find a Rosetta Stone with the Voynichese writing on it, this line of thought too remains inconclusive. Still another theory suggests that the Voynichese

is a constructed language. Many ancient languages were constructed in an attempt to develop a universal language, one that made information more accessible. But this theory contradicts the theory of the cipher and its purpose of hiding information.

Other theories on the meanings of the Voynich manuscript and its origins abound, but as long as the text remains illegible, and is in the context of this paper, the Voynich fits as a perfect example of early asemiosis. Perhaps I can go a step further and entertain the idea of it being an eco-asemic work with its treatise on nature, drawing its universal script from nature itself: floating without a human author or human meaning.

## 6. Remembering

I recently attended “Nanocosmic Investigations—Artists in Conversation with ESS” an artist residency at Inter Arts Center in Malmö, Sweden. The residency was a collaboration between Malmö Museer, The European Spallation Source (ESS) and Inter Arts Center at Lund University. ESS was building a proton accelerator and the discussions with the ESS scientists helped me understand the different forces that were exerted and controlled in order to focus and accelerate the beam of protons. What really stayed with me at the end of the discussions was the idea of a horizontal path, a horizontal travel, and all the efforts needed to make it happen. That transverse magnetism to the horizontal made me think of the positive sign ‘+’ which has both directions, the vertical and horizontal. It is also the symbol that the proton carries. I was inspired to explore ways that the vertical could go into horizontal and eventually worked out a code to visualize it. The outcome turned out to look like an active asemic script, as you can see from a still of the animation in the top section of Fig. 6.

I was then interested to explore the different ways our bodies could go horizontal while we imagined ourselves preparing for a horizontal travel along the beam (Fig. 6). For us humans, we are very familiar with the horizontal. We have evolved from it to stand upright on our legs. Yet we still go back to it when we rest and sleep or pass away. Horizontal is home. Even our text and writing are linearly horizontal in reminiscence to our original reference of home. So perhaps through that accelerator we are traveling home and through our asemiosis we are writing home. We are perhaps writing home and traveling in time while being still—at home.

I think asemic writing implies energy. As Tim Gaze puts it: “Asemic writing is a visual stimulus.” Devoid of words, it directs us towards the physicality of the trace casting light on the primal desires, feelings or energies that precede thought. Asemic writing is always active, even in its

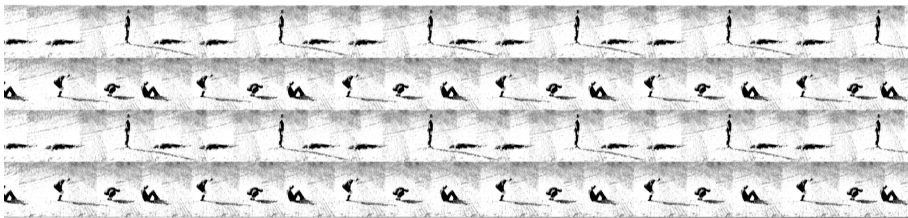
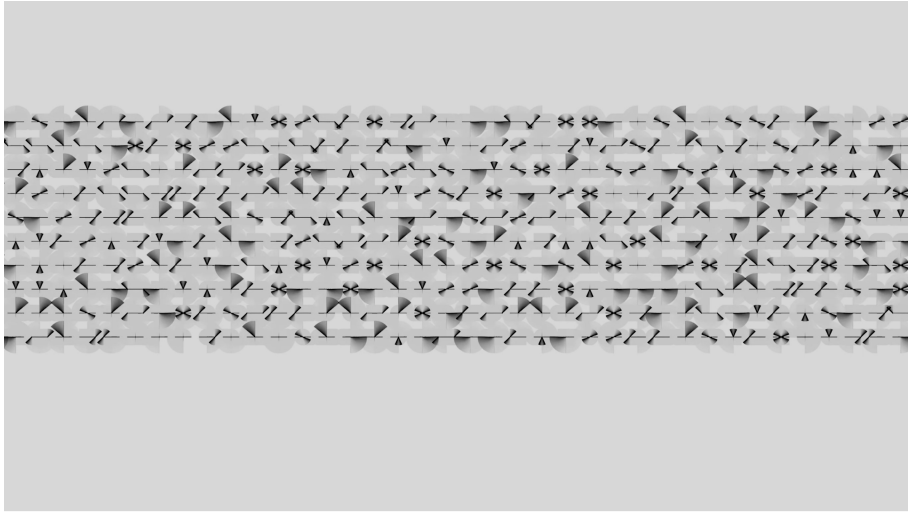


FIGURE 6. Stills from the video “Transverse” (2022) by Christine Kettaneh

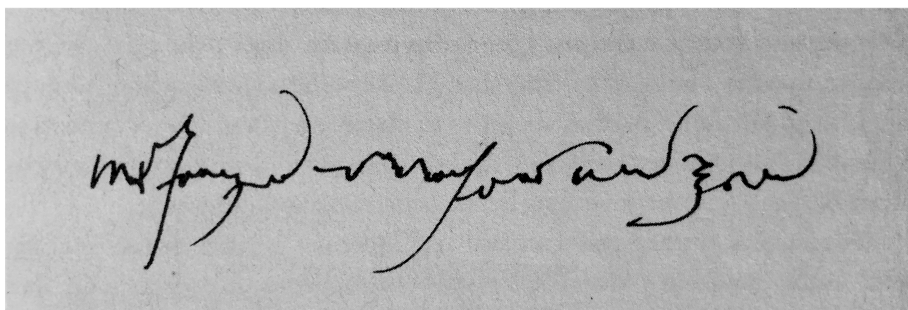


FIGURE 7. Roland Barthes, “Contre-écriture.” (Schwenger, 2019, 33, Fig. 2.5)

static form, either triggering our compulsions for meaning or inviting us to play, learn or solve. Moreover, asemic writing creates the opportunity to question: what is writing? And what is reading? Trying to answer these questions will take us ultimately back to where it all started, before the seme, before the meaning, to the flat ground that holds all other forms on it, below it or above it. As Schwenger adequately puts it: Asemic writing “may be without meaning; but it is not without significance.”

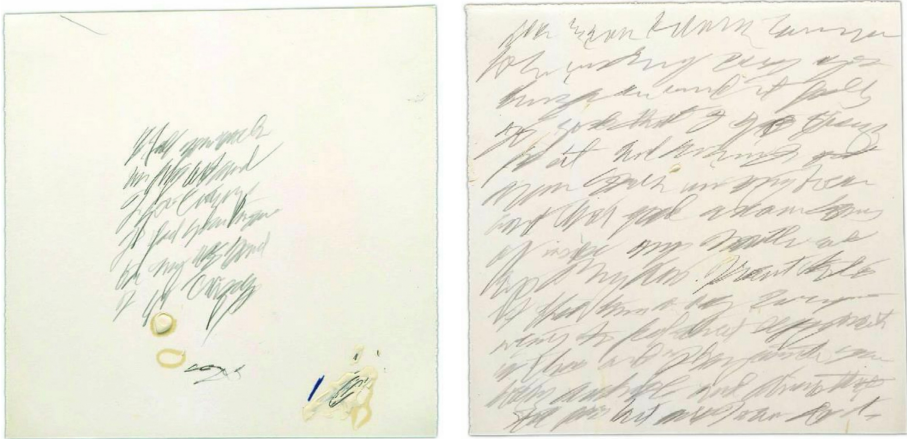


FIGURE 8. Cy Twombly, “Letter of Resignation XXV” & “Letter of Resignation XXXVI,” 1967. Copyright Cy Twombly Foundation

## References

- Babkina, Katerina (2015). “Luigi Serafini on How and Why He Created an Encyclopedia of an Imaginary World.” *Bird in Flight*. <https://birdinflight.com/media/luigi-serafini-on-how-and-why-he-created-an-encyclopedia-of-an-imaginary-world.html>.
- Badmington, Neil (2008). “The “Inkredible” Roland Barthes.” In: *Paragraph* 31, pp. 84–94.
- Barthes, Roland (1976). “Contre-écritures.” In: *Graphies*. Ed. by Marc Dachy. Vol. 2. Luna-Park. Brussels: Transédition.
- Borges, Jorge Luis (1961). “Tlön, Uqbar, Orbis Tertius.” In: *New World Writing*.
- Brownie, Barbara (2014a). “A New History of Temporal Typography. Towards fluid letterforms.” In: *Journal of Design History* 27.2, pp. 167–181.

- (2014b). "Alien Scripts: Pseudo-Writing and Asemic Writing in Comics and Graphic Novels." In: *3rd Global Conference: The Graphic Novel, Oxford, UK, 3-5 September, 2014*, pp. 1–9.
- (2015). *Transforming Type. New directions in kinetic typography*. London: Bloomsbury.
- 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, pp. 117–139.
- Comerford, Colleen (n.d.). "Book and Interactive Website." <http://www.colleencomerford.com/bookinteractive.html>.
- Ellis, Colleen (2010). *ABCing. Seeing the Alphabet Differently*. New York: Mark Betty Publisher.
- European Spallation Source (2022). "ESS art residency programme. Nanocosmic Investigations exhibition." <https://europeanspallationsource.se/article/2022/04/06/ess-art-residency-programme-nanocosmic-investigations-exhibition>.
- Gaze, Tim (2011). "Asemic Movement 1." <https://issuu.com/eexxiitt/docs/asemicmovement1>.
- (2015). "Gaze. A few persistent thoughts about asemic writing." <https://www.utsanga.it/gaze-a-few-persistent-thoughts-about-asemic-writing/>.
- Hochelaga (2022). "Secrets of the Voynich Manuscript." <https://www.youtube.com/watch?v=csyxcbqgczo>.
- Inter Arts Center (2022). "Nanocosmic Investigations. The Exhibition." <https://www.iac.lu.se/projects/nanocosmic-investigations/>.
- Jacobson, Michael (2022a). "A letter from Cecil Touchon to Peter Schwenger discussing asemic reading!" The New Post-literate. A Gallery of Asemic Writing, <http://thenewpostliterate.blogspot.com/2020/01/a-letter-from-cecil-touchon-to-peter.html?m=1>.
- (2022b). "On Asemic Writing." In: *Asymptote*.
- (2022c). "The New Post-literate. A Gallery of Asemic Writing." <http://thenewpostliterate.blogspot.com/>.
- Kettaneh, Christine (2018). "The Hindwing." <http://www.christinekettaneh.com/#/new-page-3/>.
- (n.d.[a]). "Limits of H." <https://vimeo.com/673865141/2d0cec2859>.
- (n.d.[b]). "Limits of ب." <https://vimeo.com/673865727/3fb8341de5>.
- Onnen, Serge (2008). *Drawings on Writing*. Atlanta, New York: J&L Books.
- Peck, Robert M. (2022). "Asemic Writing From the Mouth of a Snail." In: *Natural History* 130.
- Schwenger, Peter (2019). *Asemic. The art of writing*. Minneapolis: University of Minnesota Press.
- Waber, Dan (1999). "Argument." <https://vispo.com/guests/DanWaber/argument.html>.





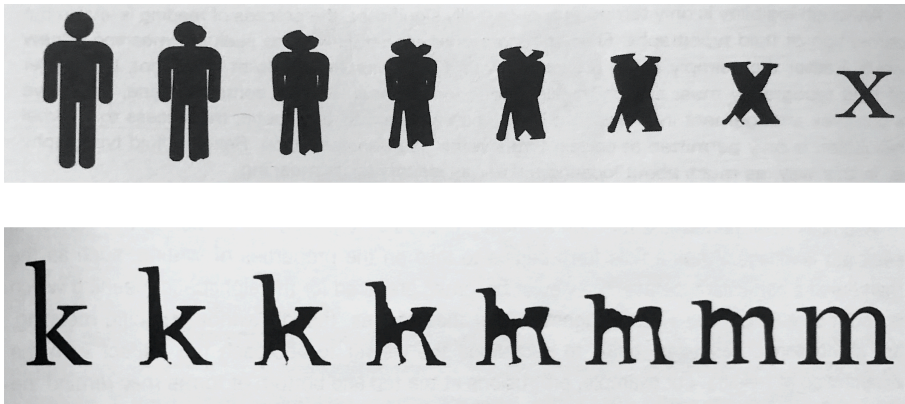


FIGURE 11. Top: “A figure of a man morphs into an “x.” During the transformation, the form evolves into an amorphous shape before becoming identifiable as a letter. The shape appears to become meaningful before its precise meaning can be discerned. During this process, the viewer must cease to perceive the image as an image, and begin to read it as a letter. Both the “x” and the man are bound up in the same form, but revealed over time. The temporal connection between these two signs is also meaningful, as it prompts the viewer not to consider each message in isolation.” © Barbara Brownie (Brownie, 2015, 52, Fig. 5.1). Bottom: “A “k” morphs into an “m.” As it transforms, the “k” ceases to be recognizable, and becomes an abstract glyph, before it eventually resolved into an “m.” At the midpoint, it is identifiable as a linguistic form of some kind, but its precise alphabetic value cannot be determined. It is at this point that it is “asemic.” © Barbara Brownie (Brownie, 2015, 53, Fig. 5.2).

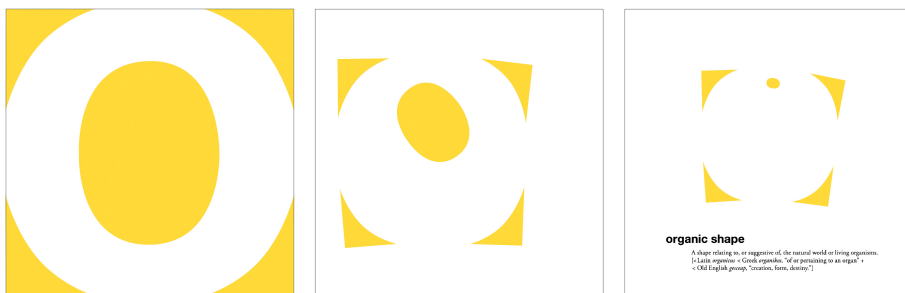


FIGURE 12. Stills from Colleen Comerford’s animation of “O” in ABCing: Seeing the Alphabet Differently (2010). The stills show the first and final poles of the transformation along with an intermediate glyph. © Colleen Comerford



FIGURE 13. Stills from Dan Waber's animation "Argument" (2005) showing a string that alternates between a "yes" and a "no." © Dan Waber



FIGURE 14. Henri Michaux, from *Mouvements*, 1951/1982. Copyright Éditions Gallimard. (Schwenger, 2019, 26, Fig. 2.3)

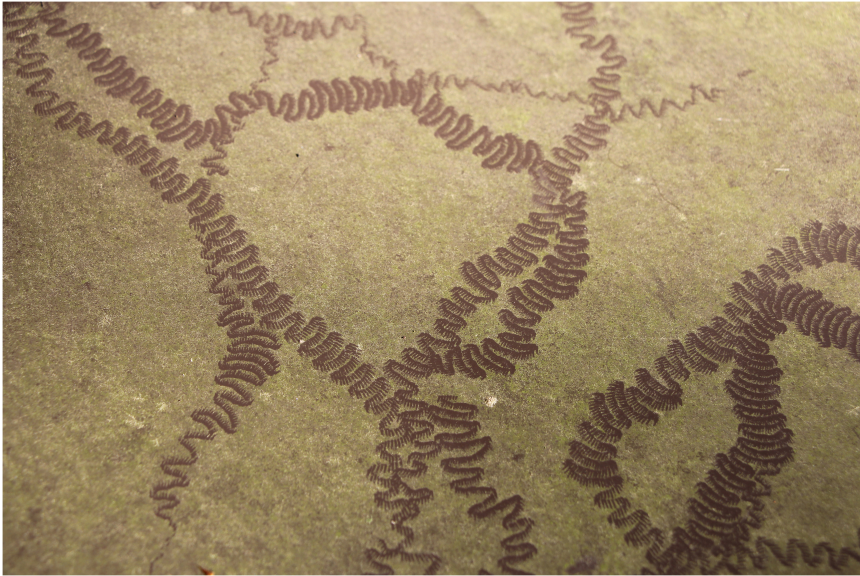


FIGURE 15. Feeding trails of a common land snail. Photo by Robert M. Peck



FIGURE 16. Trails left by a bark beetle infestation under the barks of a pine tree. Photo by Christine Kettaneh



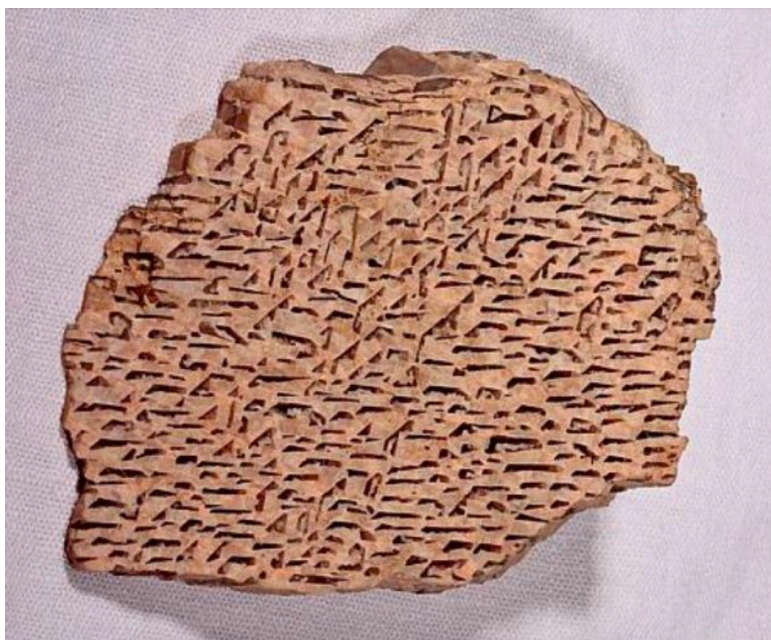


FIGURE 17. Photograph of graphic granite (NMNH 111123–1767) by Ken Larsen. Courtesy of the Smithsonian Institution (Schwenger, 2019, 65, Fig. 3.4)

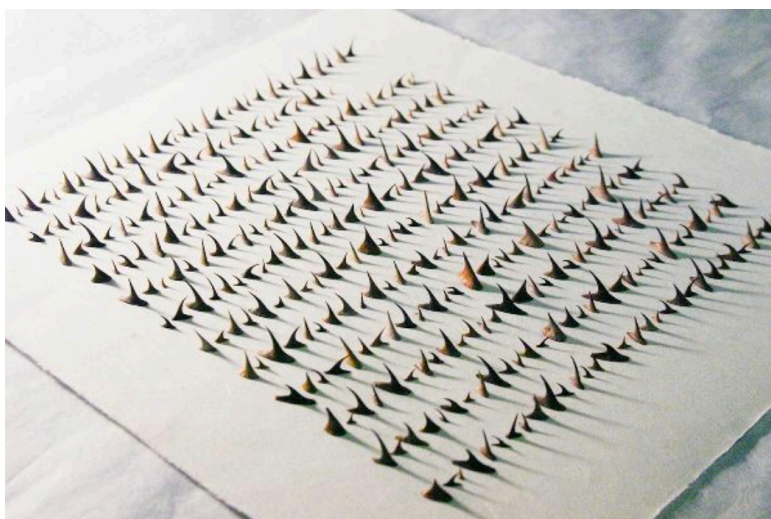


FIGURE 18. Cue Fei, "Read by Touch," 2005-6. Thorns on rice paper. Each page  $9\frac{1}{4} \times 10\frac{3}{4}$  inches; total 11 pages. Photograph by Zheng Lianjie (Schwenger, 2019, 76, Fig. 3.9)

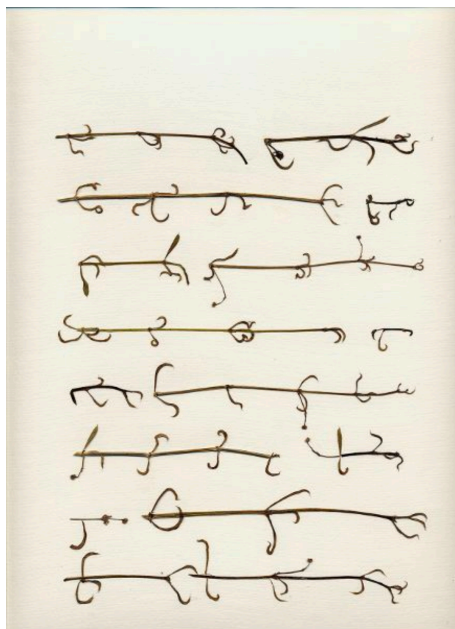


FIGURE 19. Marian Bijlenga, page from the book "Written Weed," no21, 2004. Catchweed on paper. Photograph by Marian Bijlenga (Schwenger, 2019, 77, Fig. 3.10)



FIGURE 20. Screenshot from "Asemic Writing in the Woods" (2011) by E.Samigulina/ Tae Ateh and Karen Kamak/ Yuli Ilyschanka (Schwenger, 2019, 80, Fig. 3.12)

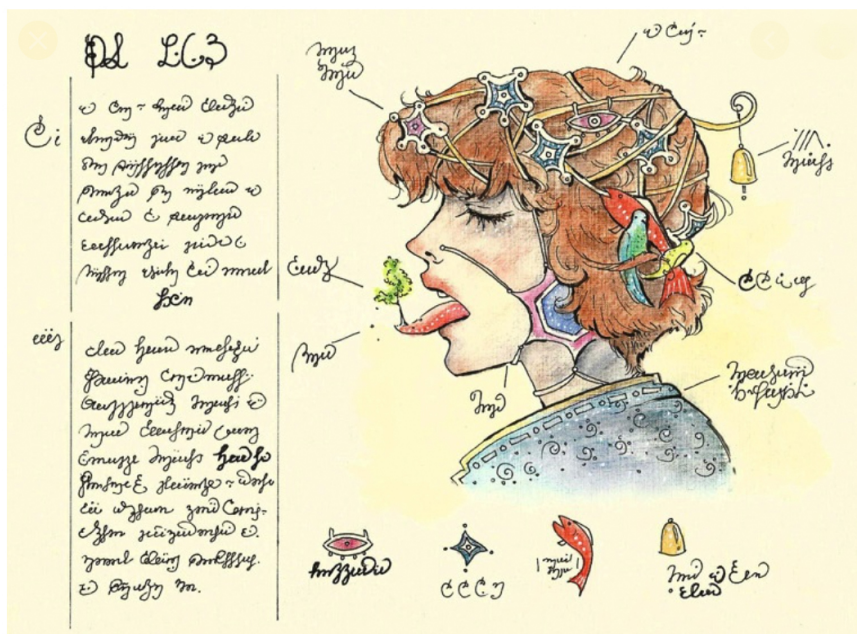
FIGURE 21. A page from Luigi Serafini's *Codex Seraphinianus*

FIGURE 22. A sample from The Voynich Manuscript



# Sinograms on Commercial Signs. A Case Study of Chinese Restaurants in Prague

Tereza Slaměňíková

*Abstract.* This paper is a response to the noticeable use of Chinese script in the public space of a country with a relatively small number of Chinese immigrants. The appearance of this linguistic phenomenon in Czechia arises from its significant involvement in the gastronomy business associated with vibrant outdoor branding. Sinograms are one of the most favorite items through which Chinese restaurant label the ethnic origin of the offered food. This study draws on the conceptual framework of linguistic landscape theory and, through visual analysis, reconstructs the graphic and linguistic contexts in which sinograms are displayed on restaurant storefronts. It is based on the photo documentation collected during August 2020 in the capital city of Prague. The sinogram-oriented approach enables a unique outlook on the dynamics of the foreign non-Latin script displayed on commercial signs. The established set of similarities indicates a high level of unity in the marketing of Chineseness through the sinograms.

## 1. Introduction

Despite the relatively short history of Chinese immigration, Chinese food ranks among the most popular ethnic cuisines in the Czech business establishments serving food. Since running a commercial property is usually associated with outdoor marketing, Chinese restaurants also contribute to shaping public space. The previous onomastic research has, among other things, revealed that the names displayed on the restaurants' outdoor signage communicate with the local consumers through three different languages and two scripts (Slaměňíková, 2023).

---

Acknowledgments: Publication of this paper was made possible with the support of the IGA\_FF\_2022\_060 Voices on the Margins of Contemporary Asia at the Faculty of Arts, Palacký University in Olomouc.

---

Tereza Slaměňíková  0000-0001-6929-7568

Department of Asian Studies, Faculty of Arts, Palacký University Olomouc, Katedra asijských studií FF UP, tř. Svobody 26, 779 00 Olomouc, Czech Republic  
E-mail: tereza.slamenikova@upol.cz

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 135–156. <https://doi.org/10.36824/2022-graf-slam>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

Apart from the Czech and English texts, the official Romanization system Pinyin for the Chinese language is widely popular. In addition, sinograms, i.e., the graphemes of the Chinese writing system, often accompany the Latin script names. Their use is so extensive that despite the fact that Chinese restaurant signs are undoubtedly not the only venue of sinograms in the public space, they are, with high probability, the most common ones. Moreover, despite a lack of statistical data, it appears to be safe to say that sinograms in all probability represented the most common non-Latin script used in the Czech linguistic landscape (LL) at the point of the data collection.

To understand its significance properly, the characteristic features of the Chinese community in Czechia have to be considered. The most recent comprehensive study emphasizes its uniqueness by comparing similar frameworks in Southern and Western Europe, as well as elsewhere in the world (Horálek et al., 2017, pp. 277ff). The attributes assigned to the Czech Chinese community include: being small and young in age<sup>1</sup> (Sluka et al., 2018, p. 89), diverse despite the relatively compact place of origin, evolving in terms of its internal composition (Moore et al., 2001) and geographically dispersed and not very communal (Horálek et al., 2017). The Chinese community in Czechia, however, shares with other countries the reality that the restaurant business is one of the main sectors of activity for Chinese immigrants (*ibid*). Therefore, although the Chinese community is small, many of these establishments in Czechia can be found. Moore et al. (2001, p. 618) mention that the number of Chinese restaurants in Prague increased from one in 1988 to almost forty in 1994. In addition, a shift in the typology of restaurants can also be observed. While Bakešová (1996, p. 364) responded in the mid-1990s to the objection as to why Czech Chinese restaurants rank among the expensive establishments, Horálek et al. (2017, p. 269) two decades later wrote that most of the several hundred restaurants run throughout Czechia are 'low-cost' restaurants offering dishes that have been altered to suit Czech tastes.

Apart from the small Chinese community, proficiency in the Chinese language in Czechia is limited to a relatively small number of sinologists and other specialists in Chinese studies, graduates of Chinese language courses with varying levels of expertise, and probably also some Chinese

---

1. Former Czechoslovakia opened its border for Chinese migration only in the 1990s (Obuchová, 2002, p. 9). The Chinese flow after the fall of the Iron Curtain came about within the so-called 'new wave' of Chinese international migration that began in the late 1980s (cf. Liu, 2005). Statistics indicate that Europe, in particular, has become an increasingly attractive destination for Chinese immigrants (Latham et al., 2013, p. 18). Although Czechia accounts for only a tiny part of the total, the growth rate of the Chinese population has increased significantly. Horálek et al. (2017, p. 269) identify the period between 1991 and 1995 as the Chinese boom during which the number of Chinese rose sixteen times from 261 to 4,210.



culture enthusiasts. Because of this, the high frequency with which sinograms appear in the public space is a phenomenon that attracts academic interest. This study aims to describe the status of sinograms in the Czech public space. Leaning on the theoretical background of LL studies, this paper documents the pragmatic functions of sinograms emerging from the context in which they are utilized.<sup>2</sup>

## 2. Theoretical Background and Research Approach

Displaying signs on the premise's front area is essential to the restaurant's branding. It is driven by the interest in directing passers-by's attention to a conducted business. To succeed in a highly competitive environment, restaurants are forced to select devices that catch the eyes of passers-by and, ideally, attract them enough to decide to become customers. Ben-Rafael (2009: 44ff) defines two major structuration principles that constitute the linguistic landscape in the central urban areas. Although they are the opposite of each other, the social actors who participate in the formation of LL are always, in a certain way, bound to adopt both of the strategies pertaining to these principles. The principle of "presentation-of-self" refers to a situation when numberless actors seek new original ways of promoting themselves and try to establish a unique signature that palpably distinguishes them from the other LL actors. This tendency grows stronger in areas with a higher density of LL items that even more notably inspire the use of unexpected devices. The "good-reasons" principle emerges from the same situation. Since actors address the same group of potential clients, they also cannot avoid adjusting their promotion techniques, including designing LL items, to align with people's expectations, values, or tastes. To achieve this, they may be induced to utilize cultural codes perceived as fashionable in the public eye or to present favorable images of themselves to others.

Language choice is one of the procedural steps underlying the preparation of any LL item. It is also one of the utterances that have the capacity to be used by the actors in favor of both of the above-mentioned principles. Spolsky (2009, pp. 34ff) describes the use of language on advertising signs as "a fine interplay" between the so-called "presumed reader's condition" and the "symbolic value condition." The rule, which the first of them is based on, states that one should have a preference for a language the presumed readers can understand to accomplish the communicative goals. The second condition emerges from the rule that one should select his or her own language or the language with which

---

2. The author of the paper would like to express gratitude to the anonymous reviewers of the original conference paper proposal. Their thoughtful suggestions were helpful in specifying the study objectives.

one wants to be identified. In this case, the choice of a particular language is motivated by the aim of evoking a specific association. Spolsky points out that advertisers apply both of these strategies when designing advertising signs. Edelman (2009, pp. 142ff) describes the same phenomenon in the way that she distinguishes two reasons standing behind the use of a particular language, i.e., transmitting factual information and appealing to people's emotions through the connotational value of languages. Edelman adopts the term "impersonal multilingualism," established by H. Haarman (1986), to refer to contexts when foreign languages are used in favor of the second reason. Cook (2013) highlights that a foreign language lends the place a certain ambiance while preferring the term "atmospheric multilingualism."

The language situation in Czechia allows for presuming that sinograms on Chinese restaurants in Prague serve a symbolic rather than a communicative function. Sinograms do not 'index' the Chinese-speaking community within which they are used: it is not their geographical placement that makes their meaning; their meaning is made by "representing something else" (Scollon et al., 2003, p. 133). Through their use, restaurants evoke an image of a different world and, thus, assert the exotic style of their cooking. According to Haarmann (1986, p. 109), "[l]anguage is the most immediate element of ethnic identity for ordinary people." Since graphic representation is an inherent part of any written language, sinograms can be seen as a direct embodiment of Chinese culture, building a link connecting food practice with ethnic identity in this particular area of usage. The question that arises is how specifically the symbolic value of sinograms on Chinese restaurant signage is constructed.

What is apparent at first glance is that sinograms are not the only linguistic code constituting the image of how the restaurants present themselves in the immediate public space. Moreover, they are rarely the only linguistic item displayed on a single sign. It therefore would seem essential to establish how they are integrated into an aggregation of linguistic systems on the storefronts. Considering the main topics investigated within the field of study on multilingual discourses in public space, two fundamental issues deserve attention. The first addresses the reader-oriented arrangement between different languages and the range of information each provides. Reh (2004, pp. 8ff) distinguishes between four types of multilingual writing: a) duplicating writing provides all the information in all languages; b) fragmentary writing displays a partial translation of the full text in one language; c) overlapping writing repeats only one part of the text in more languages, while the other parts are provided in one language only; d) complementary writing provides different information in each language. Simply speaking, any sinogram (or even a graphic unit resembling a sinogram) can create the desired illusion, yet, they are not nonsensical or randomly chosen. Thus, despite

the primarily symbolic function, the denotational meaning is a significant aspect of the usage of sinograms that needs to be discussed. In addition, an essential part of it is the typology of the ideas that are chosen to be transmitted or, in contrast, kept hidden from the local community. The second issue concerns the visual treatment of sinograms in relation to other linguistic codes. Although the visual weight of elements in a graphic composition is not objectively measurable, the information value of each of them results from their mutual interaction, which can be determined by various factors, such as placement in the composition, size, or color contrast (Kress et al., 2006, pp. 201ff).

In light of the two above-mentioned principles of LL structuration, Ben-Rafael (2009, p. 50) argues that food and restaurant establishments mainly target the recurrent needs of the local clientele and are therefore more likely to leverage cultural branding strategies that respond to the good-reasons principle. In light of this, this paper searches for repetitive patterns that might indicate a sociocultural unity in the visual communication of sinograms. It has also been observed that the global marketization of ethnicity and commodification of culture seems to have significantly impacted the development of LL in urban environments since the late twentieth century (Leeman et al., 2010a). This is also the case with Chinese ethnicity, which is often used as a marketable resource promoting exotic potential. According to Ang (2016, p. 261), “Chineseness became an object of commodification, which is often self-commodification” in western Chinatowns. Given this reality, the present paper also sheds light on the mechanisms through which one of the primary Chinese identity markers is commodified for marketing purposes.

### 3. Research Corpus and Methodology

The data collection was undertaken in the capital of Czechia. It was a reasonable choice because migrants mainly chose Prague as the place to settle down during the Chinese boom of the 1990s (Moore et al., 2001, p. 614). Data are composed of Chinese restaurants located in all ten districts of Prague. As mentioned in the introduction, the photo documentation of the Chinese restaurant exterior was initially taken to examine the restaurant naming practices.<sup>3</sup> In its processing, the vast popular-

---

3. The author of this text would like to express her appreciation to two students from the Department of Asian Studies, Palacký University in Olomouc, namely Mgr. Michaela Frydrychová and Bc. Terezie Kadlecová, for collecting this photo documentation in August 2020. The different purpose of their collection was not associated with a high demand on quality. For this reason, the photos used in the figures in this paper were retaken in September 2022 by the author.

ity of sinograms displayed on the outdoor signary inspired a new approach to the data set that targets the foreign script elements shown in the restaurant signage. A sample of 120 Chinese restaurants, displaying at least one sign with sinograms, was examined. Restaurants located in shopping malls, pedestrian underpasses, and passageways through a building were not included in the analysis in cases when their front window was not visible from the street view.

To achieve its goals, this paper employs methods of visual analysis that represent an inherent part of the LL study (cf. Scollon et al., 2003). It should be pointed out, however, that the adopted approach deviates from its traditional sphere of interest which usually embraces the full spectrum of linguistic items displayed on a geographically coherent whole. In contrast, this study focuses on one specific constituent observed on one particular segment of objects in many separate locations. It is also not motivated by the traditional aim of measuring linguistic diversity in multilingual contexts (cf. Landry et al., 1997). It aspires to establish the status of one specific foreign linguistic phenomenon in an essentially monolingual country. Applying the primary classification designed by Barni et al. (2009), the analyzed segment is characterized by an external position and location in both central and peripheral urban areas; it belongs to the public domain and its subcategory catering. Taking into account Scollon et al.'s (2180ff) categorization of texts in urban spaces, the discussed constituent is displayed on commercial signs. These signs are private in terms of authorship (cf. Landry et al., 1997) and bottom-up in terms of the source they stem from (cf. Ben-Rafael et al., 2006).

Using a quantitative approach, this study first isolates the occurrences of the foreign script elements on the analyzed segment and, second, focuses on the most frequent sources of sinograms and describes the linguistic and graphic context in which they operate. It attempts to identify the recurrent strategies developed in constructing the symbolic function of sinograms. After reviewing the collected photo documentation, a set of three research perspectives was established to classify the data: 1) placement of the sinograms on the bounded physical space of the restaurant fronts; 2) semantic content of the writing in sinograms and its relationship to writings in other languages; 3) graphic presentation of sinograms within multilingual signage. The first perspective perceives the restaurant storefront as the research unit. After reviewing the spectrum of semantic contents, the second perspective transfers its attention to the arrangement of multilingual writing on a single sign. The same object is being targeted while approaching the data from the third perspective. The numbers in the brackets, used throughout the paper, indicate the total amount of the currently described facts.

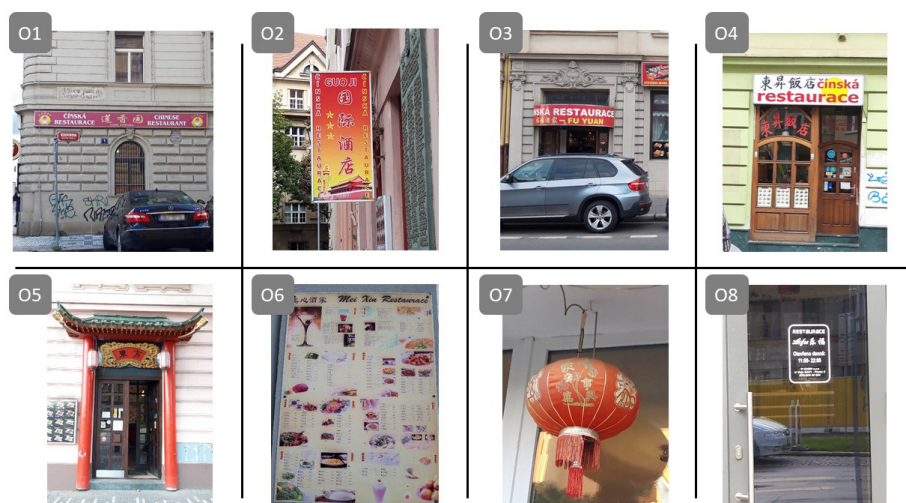


FIGURE 1. Objects displaying sinograms

## 4. Findings

### 4.1. Sinograms and Emplacement

Chinese restaurants in Prague are not concentrated in one specific location but are dispersed all around the city. Apart from several kilometers-long avenues, they are rarely placed on the same street. All the analyzed restaurants use a complex set of visual formats on their storefronts through which they differentiate themselves from the surroundings, including physical manifestations of language such as sinograms. This section explores the location of sinograms in the composition of signage. The list below summarizes the types of objects where sinograms are displayed. Examples of each of them are provided in Figure 1.

- O1. Upper wall signs: plates in a shape of a longer dimension horizontal rectangular fastened parallel to the wall and located in the upper part of the building's first floor.
- O2. Overhanging signs: signs attached to a building in a way that projects over the street.
- O3. Other large-size signs: a) signs of different formats attached to other parts of the building; b) mounted sinograms attached to a wall surface or erected over the top of the roof; c) permanent freestanding signs.
- O4. Window signs: sinograms painted on the windows.
- O5. Chinese architectural elements: sinograms displayed on an entry in the form of a traditional Chinese gate.

## O6. Menus.

O7. Chinese decorative artifacts: sinograms written on lanterns hanging outside the building. Apart from this, the sinograms are sometimes on other Chinese artifacts (vases, statues, lanterns, knot decorations) located in the restaurants' interiors, which are also clearly visible through the window.

O8. The statement of the place of business<sup>4</sup>.

Generally, the amount of signage displayed on any commercial premise is influenced by its location. Chinese restaurants in corner buildings often place, for example, the same signs on both sidewalls. Restaurants in freestanding buildings sometimes even exploit the potential of placing the sign on the roof. Most of the analyzed restaurants run their business, however, in a building closely surrounded by other buildings in the street. Thus, using multiple signage is simply one of their promoting techniques. In a few cases, the front of the building is separated by a front yard from the street, and signage is also located on the front yard fence.

Table 1 shows how many signs with sinograms can be found in an individual restaurant. It is centered on the objects visible from the street vantage point, not the small items recognizable only in close proximity to a restaurant. This parameter applies to objects provided under the numbers O1 to O5 in the above list. As can be seen, restaurants often display sinograms on at least two signs. The most productive is a combination of an upper wall and an overhanging sign that applies to both groups with multiple signage.

TABLE 1. Number of signs

| Number of signs | Total | Types of signs           | Total |
|-----------------|-------|--------------------------|-------|
| One             | 31    | O1                       | 11    |
|                 |       | O2                       | 13    |
|                 |       | Other signs              | 7     |
| Two             | 50    | O1 + O2                  | 24    |
|                 |       | Other combinations       | 26    |
| Three or more   | 39    | O1 + O2 + any other sign | 16    |
|                 |       | Other combinations       | 23    |

4. According to the Czech Trade Act, every establishment intended for provision of services to customers must be permanently and visibly marked from outside. The law does not specify what a sign should specifically look like. It is quite common to mark the place of business on the entrance door. Some of the Chinese restaurants provide the restaurant name in sinograms as well.

## 4.2. Sinograms and Information Arrangement

Sinograms are rarely the only unit displayed on a single sign. Most are implemented in a complex composition of items in different linguistic codes. This section explores the interaction of sinograms with these codes. First, it classifies the meaning that is communicated through sinograms. Second, it explores to what extent the message in sinograms is shared by means understandable by the local community and characterizes the features of not transmitted ideas. Third, it evaluates the role of sinograms from the perspective of Reh's (2004) arrangement of multilingual writing.

Sinograms displayed on storefronts are not randomly chosen graphemes. They usually transmit complex pieces of information. These messages can be divided into five groups:

### M1. Name (111).

This type of message can be found on all the objects listed above. Generally speaking, a restaurant name can be composed of two parts. An obligatory part, the so-called specifics, identifies the particular commercial establishment. A facultative part, the so-called generics, refers to a general class of names, i.e., a place where meals are prepared and served to customers. Only about half of the restaurant names in sinograms (58) in the analyzed sample contain both the specifics and the generics. The generics in sinograms are represented by different Chinese words expressing the concept of a restaurant, i.e., 饭店 (36), 酒家 (7), 酒店 (5), 酒楼 (3) and 食府 (1), or specifying the sort of offered dishes, i.e., 快餐 (5) 'fast food' and 美食 (1) 'delicious food.' The specific parts of the names proceeding the generics are semantically heterogeneous. In general, the choice of lexical units follows the recommended strategies for the restaurant or commercial names described in different Chinese handbooks for name creation (e.g., Chen et al., 2011, pp. 279ff; Dong, 2012, pp. 193ff; Mao et al., 2003, pp. 94ff). The names also demonstrate similarities with the tendencies observed for brand names regarding the importance of positive connotations (cf. Basciano, 2017; Chan et al., 1997; Chan et al., 2001; Chan et al., 2009).

### M2. Names or types of dishes (14).

The occurrence of this type of message is limited to O6 menus. The different approach to fixed-price meals (including side dishes, usually served during lunchtime) and non-fixed food items (offered all day long) is of interest. The use of sinograms is limited to the latter. Some menus only attach sinograms to selected meals or general categories.

### M3. Type of business (8).

The text in sinograms provides a hint about the type of business (which is not part of the name), e.g., 中餐厅 'Chinese restaurant,' 川

菜 ‘Sichuan cuisine,’ or a two-line text 中式佳肴 ‘Chinese delicacies’ (first line) and 家的味道 ‘home-style flavor’ (second line). This type of message was observed on objects O1, O2, and O3.

M4. Wishes for prosperity and good fortune (5).

The desire for auspiciousness pervades many levels of everyday life in Chinese culture, including business activities. A practical way to secure its steady flow is to display auspicious symbols, such as the sinogram 福 ‘good fortune.’ Another widespread practice is based on materializing relevant sayings while writing them down, e.g., a four-sinograms structure 恭喜發財 ‘May you be happy and prosperous.’ Locations for this type of message include objects O1, O4, O5, and O7. The total number provided in the category headline takes account of objects displayed in the restaurant exteriors. Interior objects with a different visibility through the window are not included.

M5. Other (3).

This group includes decorative elements in restaurant logos other than restaurant names, e.g., the sinogram 味 ‘taste’ over the steam rising from a bowl.

Since Chinese is a language primarily unfamiliar to the local community, the question arises as to what extent the messages in sinograms are communicated in a language Czechs can understand. Messages of M4 and M5 are provided only in sinograms. In the case of M3, the amount of transmitted information is based on the complexity of the text: simple terms referring to the type of business are usually also provided in Czech or English; more detailed descriptions only appear in sinograms. M2 messages are mostly simultaneously offered in Czech, in many cases also in English and occasionally in German. The most common order in sinograms is sinograms—Czech—English—German. The extent to which the Latin script texts represent direct or loose translations varies among the dishes and restaurants and is a topic for a separate research paper due to its high complexity. The largest group of M1 restaurant names shows significant differences as concerns the specifics and generics which are, therefore, discussed separately.

The generics are often also a part of the Czech name (49). The Czech nomenclature is not as developed, however, as in sinograms. It is limited to the Czech version of the international term restaurant and two terms capturing the quick-service concept. It is also not unusual that the Czech generics is supplemented or even replaced by an English one. What has to be pointed out is that the Czech or English generics are repeatedly extended by the attribute ‘Chinese’ or ‘China.’ The same construction is often part of the Czech name even when the generics are not included in sinograms. This practice can be seen as additional evidence that designing signs is driven by the aim to clearly mark the origin of dishes.



Contrary to the generics, the message hidden in the specifics is rarely transmitted in Czech. The reason is that the name-givers prefer the official Romanization system Pinyin while providing the name in Latin script. Table 2 summarizes the practices as to how the specifics are transliterated. Proper names of geographical origin are displayed separately. This is because it is impossible to draw a strict line between names of well-known destinations and names that may not be recognized as Chinese toponyms by Czechs. As can also be seen, the Czech or English specifics are not necessarily word-by-word identical to the name in sinograms. Finally, five restaurants display the specific part of their name only in sinograms without offering their Latin script version.

TABLE 2. Latin Script versions of the specifics in sinograms

| Category | Total | Subcategory      | Total | Examples   |
|----------|-------|------------------|-------|--|
| Pinyin   | 67    | Fully identical  | 67    | 福达 <i>Fu Da</i>                                      |
| Czech    | 15    | Fully identical  | 6     | 莲花 <i>Leknín</i> ('lotus flower')                    |
|          |       | Partly identical | 9     | 红樱桃 ('red' + 'cherry') vs. <i>Třešeň</i> ('cherry')  |
| English  | 7     | Fully identical  | 4     | 阳光 <i>Sunshine</i>                                   |
|          |       | Partly identical | 3     | 明月楼 ('bright' + 'moon' + 'building') vs. <i>Moon</i> |
| Toponym  | 17    | Pinyin           | 13    | 扬子江 <i>Yang Zi Jiang</i>                             |
|          |       | Other            | 4     | 四川 vs. <i>S'chuan</i> (non-standard transliteration) |
| None     | 5     | None             | 5     | 悠悠阁  |

In sinograms, most specifics are refined combinations of carefully chosen linguistic units that evoke culturally grounded positive connotations. They are expressed through explicit references to good fortune, prosperity, and enjoyment or culturally shared auspicious symbols, especially plants and precious substances. Name-givers also like to allocate these expectations to a particular place as a symbolic substitute for the restaurant itself. Taking into account the limited occurrence of the Czech specifics, it is clear that only a tiny portion of these motifs can be shared with the host country through the local language. Table 3 divides the used linguistic units into several semantic groups. The left side of the table lists the concepts abstracted from the Czech specifics. The right side summarizes the most common ideas transmitted only in sinograms. Names covering more semantic groups are numbered in each of the semantic groups.

The table demonstrates that the reference to a place with a certain ambiance is much more developed in sinograms. Apart from this, the

TABLE 3. Transmitted and hidden ideas

| Transmitted ideas   | Hidden ideas   |
|---|--|
| Plant motifs (7)<br>Garden (4)<br>Pleasant smell, pearl, happiness,<br>harmony, new age (1) | Positive expectations:<br>– Prosperity, abundance, wealth (16)<br>– Happiness, good fortune (14)<br>– Pleasure, joy (10)<br>Places with a certain ambiance:<br>– Garden (9), building (7), pavilion (5)<br>– Home, family (6)<br>Auspicious symbols:<br>– Precious substances (5)<br>– Animals (3) |

Czech specifics almost omit literal implications of positive expectation. Instead, they show the somewhat surprising popularity of motifs related to the world of flora where one would not expect the local audience to translate their often exceedingly manifold symbolic status in Chinese culture (cf. Slaměniková, 2023).

The previous description was sinogram-oriented. From the point of view of the arrangement of multilingual writing, the way the generics are displayed on the signs mainly matches the category of fragmentary writing. The complete information includes the derivate of the term China and is provided in Czech or English. The arrangement of the specifics encounters the problem of how to evaluate the different graphic representations of the same language. Pinyin duplicates the message provided in sinograms. Swapping sinograms for Latin script letters is merely, however, a formal adaptation that does not involve a meaning transfer. In fact, the effect is precisely the opposite. Once dissociated from the sinograms, Pinyin names become ambiguous due to the high level of homophony in Chinese. Thus, the relationship between sinograms and Pinyin cannot be considered duplicating. The exceptions are the specifics designating Chinese toponyms, in which case it is a common practice to incorporate them without translation into Czech. Finally, the arrangement between the specifics provided in sinograms and Czech/English is based on the amount of their identity, either duplicating or fragmentary.

The arrangement, however, of the specifics and the generics, as a coherent whole displayed on one sign, is more complex. One part of its structure is communicated in sinograms, and one part is in Latin script. The following schema summarizes the combinations that the unit in fo-

cus, i.e., the specifics in sinograms, create with the other items involved. The dashed line indicates a facultative item. The schema also depicts the two most common combinations. The first of them belong to overlapping writing: there is only one element shared in both Chinese and Czech/English, i.e., the type of business. As mentioned above, Pinyin is not helpful in terms of the transmission of information. In the case of the second combination, the languages complement each other.

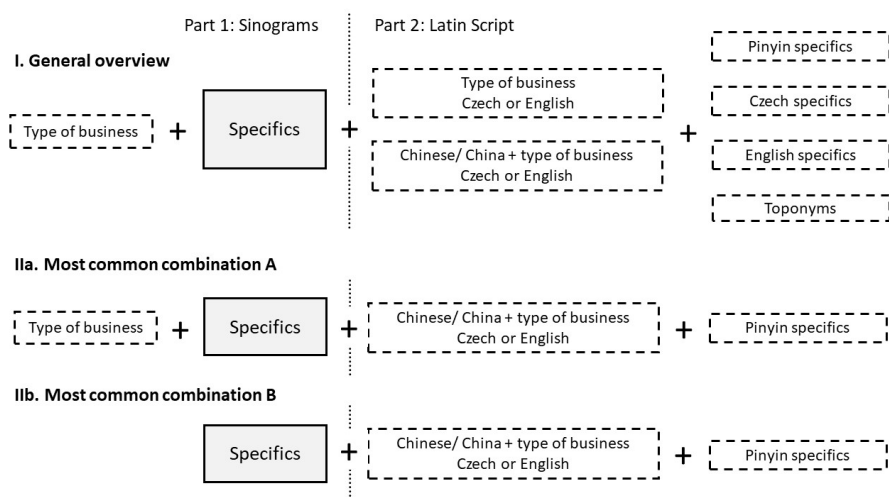


FIGURE 2. Sinogram-centered overview of script and language combinations

### 4.3. Sinograms and Graphic Design

This section explores how sinograms are communicated graphically. Targeting the O1 signs and those O3 signs situated at a height above the window or higher, it describes the spatial context of the sinograms within a particular sign and the use of typographic devices. The discussed signs have a horizontal rectangular shape, usually with straight lines, but two signs are arc-shaped. Rather than provide a comprehensive picture, this section searches for repeated patterns indicating the most common choices for displaying sinograms on storefronts.

As far as the placement of sinograms is concerned, the following four groups can be distinguished. Figure 3 provides an overview of the most common patterns, including examples of the actual signs. The graphic representations of the patterns have to be understood as simpli-



FIGURE 3. Most common layout patterns

fied schematic versions that do not attempt to depict the accurate proportions of the demarcated parts of the pattern. For practical reasons, they also refrain from capturing the potential visual aids or occasional small-size texts that merely supplement the main linguistic items of the sign.

(a) Sinograms attached to a logo (13)

LG: Logo is usually located on the left side of the sign, and the font size of sinograms is relatively tiny.

(b) Sinograms on the sign with a horizontal layout (37)

The two most common patterns were identified:

- H1 (17): The sign has a three-column layout displaying sinograms in the middle and the Latin script on the left and right sides. The Latin script texts on the sides are either complementary to each other, e.g., one of them provides the generic part and the other the specific part of the name; or they duplicate the same information in Czech and English. A variant of this pattern marked as H1b (9) includes the more minor size specifics in the Latin script placed under the larger size sinograms.
- H2 (10): The sign has a two-column layout in which one of the columns is split into an upper and lower part. Sinograms either occupy the prominent non-divided column (H2a) or are located in one of the smaller blocks in the divided column (H2b).

(c) Sinograms on the sign with a vertical layout (27)

The vertical structure shows a high degree of diversity and includes two-, three- and sometimes even more-level patterns. Only one of them appears with a higher frequency:

- V1 (10): the sign utilizes a two-level pattern with sinograms placed on the left or right side of the lower part divided into two columns. The upper part is occupied by the Czech generics that represents the most common item located on the highest level on all the signs with a vertical layout.

(d) Sinograms on other signs (11)

These signs possess unique complex layouts with horizontal and vertical levels pervading one another.

As can be seen, the organization of entities on the sign usually exhibits a hierarchical structure. It is not unusual that sinograms are displayed in a larger font and thus represent the most eye-catching item (34). This is how, for example, the entities on an H1 sign and an H2a sign are mainly implemented. Another favorite practice can be observed in the examples of H2b and V1 patterns: sinograms are less prominent, but their level in the hierarchy is identical to another entity (13). It is also quite common that neither of the items appears to be favored through its typographic qualities, especially on signs with a vertical layout. Sinograms are rarely placed on the highest level and, thus, following the code preference criteria developed by Scollon et al. (2003, pp. 116ff), do not represent the most prominent code. Despite this fact, it should be highlighted that signs with proportionally much smaller sinograms appear relatively rarely (14). In most cases, they are visually significant enough to be distinct from a physical distance.

Using typographic devices demonstrates strong preferences in color and Chinese writing style. Restaurants often choose red and yellow/gold, representing Chinese culture's lucky colors.<sup>5</sup> Red is the most common background color (47). Texts in white and yellow usually supplement it: they are used either individually for the whole text (white 11, yellow 4) or, more often, combined in a way that they somehow distinguish texts of different linguistic codes (30). The most common non-red backgrounds include white to light grey (11), black (10), and yellow (7). At least one part of the text is provided in red on most of these signs. As concerns the sinograms, they appear in five different colors in all: yellow (33), red (23), white (22), black to dark blue (9) and green (1).

In Chinese calligraphy, sinograms can be written according to several writing styles whose origin is linked to the historical development of Chinese script. The analysis of the most prominent sign of each

---

5. Red is the national color associated with good fortune and happiness; yellow/gold is the royal color that symbolizes prosperity (cf. Williams, 1976, pp. 76ff).

restaurant has revealed that Chinese restaurants in Prague prefer the graphically abbreviated semi-cursive script (73) over the modern standard regular script (44). Brush leaves the paper less often in semi-cursive writing and as a result, strokes tend to run into one another. The text is more decorative but, at the same time, less legible. The appearance of ancient clerical and seal script is limited to single units. In addition, restaurants also take into account the two versions of sinograms emerging from the script reforms in the last century. It has been observed that most restaurants prefer simplified sinograms. The use of their older pre-reform version, i.e., traditional sinograms, is limited to 11 restaurants.

Finally, it should be mentioned that the rectangular shape of signs is not suitable for vertical writing, i.e., a text vector that also enjoys overall popularity in Chinese texts. Chinese restaurants in Prague sometimes utilize this arrangement on overhanging and window signs. Another script-specific form of visual communication, i.e., right-to-left writing, is limited to two traditional gate signs.

## 5. Discussion

Sinograms are an essential part of restaurants' branding. The fact that they rarely appear in the Czech public space makes them a powerful tool through which geographically dispersed Chinese restaurants compete for visibility in the linguistic landscape. Responding to the self-presentation principle, restaurateurs place themselves in contrast with surrounding establishments. At the same time, preference for certain practices indicates that sinograms also represent a sociocultural clue operating in favor of the good-reasons principle. The examination of sinograms' interaction with other linguistic elements on different store-fronts has revealed a range of similarities that can be described as follow:

### 1. Reiteration.

Sinograms are often displayed on at least two large-size signs. Regarding the typology of signs, the occurrence of sinograms on an upper wall sign and an overhanging sign is the most significant. An essential attribute of most signs is multilingualism.

### 2. Exclusive content.

The arrangement of the multilingual writing on signs displaying the sinograms mainly varies from overlapping to complementary. The information transmitted in all languages involved is mostly limited to the generic term identifying the type of business. As has been demonstrated, however, the relationship between the elements provided in the foreign Chinese writing system and the local Latin script

is more complex. The reason for this is linked to the popular practice of how the specific parts of the restaurant names are established. Most of them are transliterated and thus represent semantically empty versions of the originally carefully chosen names in Chinese. In other words, an attempt to converge with the local community in terms of the form but not the content is apparent.

3. Typification in the visual appearance.

Signs with sinograms demonstrate a high degree of unity in graphic design. Elements provided in different linguistic codes are physically separated, and the layout of the signs decomposes into several main parts, each represented by a visually prominent linguistic entity. The analyzed rectangular-shaped signs favor two horizontal patterns and one vertical pattern. The most popular placement of the sinograms is the central position in a graphically symmetrical horizontal layout. Although not necessarily the most salient items, sinograms frequently compete for passers-by's attention through visually appealing typographic qualities. Signs with sinograms are characterized by the three most frequent colors: red, yellow and white. Finally, restaurants tend to utilize the writing style with a higher decorative effect.

These three points summarize how the symbolic value of sinograms is predominantly constructed and imply a certain unity in the marketing of Chinese ethnicity on restaurant storefronts in Czechia. In addition, some of the findings are congruent with the use of Chinese in multilingual regions. Red and yellow/gold were observed, for example, as two colors pervading the outdoor signage and menu designs of two case study restaurants in Paris Chinatowns (Lipovsky et al., 2019, p. 227). Interestingly, green as a third color, prevailing on the color schemes across Chinese-run establishments in Washington, D.C.'s Chinatown (Lou, 2007, p. 188), is almost omitted on the Czech Chinese restaurant storefronts. The author of the research in Washington D.C. also observed, however, the popularity of the horizontal symmetrical layout of the signs (ibid, p. 181). Generally speaking, its popularity can be attached to the perception of the central composition as the fundamental organizing principle in Chinese visual semiotics (cf. Kreuss, 2006, p. 195). Lou (2007, p. 181) distinguishes between two basic strategies for designing these signs, i.e., splitting the Chinese name or its repetition. They are, to a certain extent, also adopted by Czech Chinese restaurants, with, however, a significant difference. Sinograms are displayed as the central item surrounded by split Latin script texts. Using both simplified and traditional sinograms indicates that designing a sign involves choosing between two orthographies, typical for places with a long history of Chinese immigration (cf., Lou, 2007; Shang et al., 2017). The strong preference for simplified sinograms in Czechia corresponds with the fact that the Chinese immigrants who left mainland

China after the language reforms are less likely to have an emotional attachment to traditional sinograms. At first glance, this strategy might seem to support Lipovsky et al.'s (2019, p. 226) finding involving prioritizing legibility over tradition in Paris. The more vigorous vitality of the semi-cursive script over the standard script indicates, however, a desire for decorativeness over ease of reading.

Contrary to the similarities in the visual appearance, the pragmatic meaning of writings in sinograms, displayed in Chinese restaurants in Prague, is shaped with a somewhat different dynamic. One of the main issues pervading the studies conducted in areas with historically concentrated Chinese people is a shift in code preference resulting from the change in power relations over time (e.g., Leeman et al., 2009; Lou, 2007, 2010; Lipovsky et al., 2019; Shang et al., 2017; Zhang et al., 2020). The emphasis is therefore placed on the interaction of Chinese with the language of the territory surrounding the Chinese enclave or with other official languages of the region, and, to a lower or higher extent, it is driven by an interest in determining the proportion between its communicative and symbolic function. The situation in Czechia is different. The relatively small Chinese population does not cluster in close geographic spaces. Horálek et al.'s (2017, p. 269) comment that most Chinese restaurants alter the offered dishes to please Czech tastes implies their focus on Czech customers. The primarily symbolic function of sinograms does not exclude, however, their communicative role for a potential Chinese clientele. The signs are polysemous and, thus, impart different messages to different groups of viewers. In this respect, the research undertaken in areas with a high concentration of Chinese population highlights the mainly informative content of the text in sinograms for Chinese speakers (cf. Leeman et al., 2010b, p. 179; Lipovsky et al., 2019, p. 227; Lou, 2010, pp. 101ff; Shang et al., 2017, p. 195). Sinograms on the Chinese restaurants in Prague often do not express any factual information since they represent a proper name in the vast majority of its occurrences (cf. Edelman, 2009, p. 151). What seems to be more prominent is the attempt to please the Chinese clientele with the selections of favorable specifics that meet the manifold requirements of commercial name designing. Compared to European naming practices, an interesting component of the Chinese onomasticon is that, apart from designing the name with the aim of evoking a particular image of the restaurants directed to the customers, the message hidden in the name can contain a wish for good fortune and prosperity in business directed to the restaurateurs themselves. This could be one of the reasons why they choose to display them on outdoor signage. Verification of the validity of this assumption requires, however, interviews with restaurateurs. They will be conducted in the next step of this research. At this point, it can be concluded that sinograms displayed on Chinese restaurant storefronts in Prague primarily function as symbols appealing to



customers' emotions, whether they are or are not proficient in the Chinese language. The difference lies in the level of linguistic analysis that the name-givers expect from different groups of viewers while decoding what they represent. It is polarized between superficial recognition of the Chinese writing system and in-depth comprehension of multilayer cultural concepts.

## 6. Conclusion

The general orientation of LL study on urban areas with large multilingual populations leads to research topic choices that, apart from English as the current *lingua franca*, tend to overlook the appearance of a foreign language in a predominantly monolingual region. The research presented in this paper highlights the significance of a non-Latin script displayed on the commercial signs in ethnically very homogeneous Czechia. The role of sinograms on Chinese restaurant signs is investigated through a visual analysis that brings together linguistic and graphic perspectives. The paper lists the types of objects on which sinograms are displayed, classifies the kind of message written in sinograms, analyzes the typology and amount of ideas transferred into Czech, and explores the graphic attributes of the displayed sinograms. The results of the analysis provide evidence that marketing Chinese ethnicity through the sinograms manifests a high level of socio-cultural unity. Three main similarities were identified: reiteration of sinograms on multiple signs, typification in the visual appearance, and exclusive content hidden from the local consumers. The first of them demonstrates the significance of sinograms in marketing Chinese ethnicity. The second indicates a tendency to an aesthetic formalization of promoting Chineseness through sinograms. Finally, the third refers to the somewhat paradoxical fact that restaurants invest significant effort in creating semantically appealing word-formation constructions in Chinese, but refrain from uncovering these culture-determined ideas to local clientele.

## References

- Ang, Ien (2016). "At Home in Asia? Sydney's Chinatown and Australia's 'Asian Century.'" In: *International Journal of Cultural Studies* 19.3, pp. 257–269.
- Bakešová, Ivana (Nov. 1996). "'Čínský svět' v českých zemích II. ['Chinese World' in Czech Countries II.]" In: *Nový orient* 51.10, pp. 363–366.

- Barni, Monica and Carla Bagna (2009). "A Mapping Technique and Linguistic Landscape." In: *Linguistic Landscape. Expanding the Scenery*. Ed. by Elana Shohamy et al. New York and London: Routledge, pp. 126–140.
- Basciano, Bianca (2017). "Brand Names." In: *Encyclopedia of Chinese Language and Linguistics*. Ed. by Rint Sybesma et al. Leiden: Brill, pp. 311–318.
- Ben-Rafael, Eliezer (2009). "A Sociological Approach to the Study of Linguistic Landscapes." In: *Linguistic Landscape. Expanding the Scenery*. Ed. by Elana Shohamy et al. New York and London: Routledge, pp. 40–54.
- Ben-Rafael, Eliezer et al. (2006). "Linguistic Landscapes as Symbolic Construction of the Public Space. The Case of Israel." In: *Linguistic Landscape. A New Approach to Multilingualism*. Ed. by Durk Gorter. Clevedon, Buffalo, and Toronto: Multilingual Matters, pp. 7–30.
- Chan, Allan K. K. and Yue Yuan Huang (Mar. 1997). "Brand Naming in China. A Linguistic Approach." In: *Marketing Intelligence & Planning* 15.5, pp. 227–234.
- Chan, Allan K. K., Yue Yuan Huang, and X. Wu David (2009). "Chinese Brand Names and Global Brand Names. Implications from Two Corpus Analyses." <https://www.yumpu.com/en/document/view/10353091/chinese-brand-names-and-global-brand-names-implications-from->.
- Chan, Allan K. K. and Yue-Yuan Huang (2001). "Chinese Brand Naming. A Linguistic Analysis of the Names of Ten Product Categories." In: *Journal of Product & Brand Management* 10.2, pp. 103–119.
- Chen, Rongfu 陈荣赋 and Liangzhu Sun 孙良珠 (2011). 好名字好前程 [*Good Name, Good Future*]. Beijing: 新世界出版社 [Xin Shijie Chubanshe].
- Cook, Vivian (2013). "The Language of the Street." In: *Applied Linguistics Review* 4.1, pp. 43–81.
- Coulmas, Florian (2009). "Linguistic Landscaping and the Seed of the Public Sphere." In: *Linguistic Landscape. Expanding the Scenery*. Ed. by Elana Shohamy et al. New York and London: Routledge, pp. 13–24.
- Dong, Yilin 董易林 (2012). 起名开运宝典 [*The Book on Creating Names that Bring Good Luck*]. Beijing: 中国物质出版社 [Zhongguo Wuzhi Chubanshe].
- Edelman, Loulou (2009). "What's in a Name? Classification of Proper Names by Language." In: *Linguistic Landscape. Expanding the Scenery*. Ed. by Elana Shohamy et al. New York and London: Routledge, pp. 141–154.
- Haarmann, Harald (1986). "Verbal Strategies in Japanese Fashion Magazines—A Study in Impersonal Bilingualism and Ethnosymbolism." In: *International Journal of the Sociology of Language* 58, pp. 107–121.
- Horálek, Adam, Ter-hsing James Cheng, and Liyan Hu (2017). "Identity Information and Social Integration. Creating and Imagining the Chinese community in Prague, the Czech Republic." In: *Contemporary Chinese Diasporas*. Ed. by Min Zhou. Singapore: Palgrave Macmillan, pp. 263–283.

- Kress, Gunther and Theo van Leeuwen (2006). *Reading Images. The Grammar of Visual Design*. 2nd ed. London and New York: Routledge.
- Landry, Rodrigue and Richard Y. Bourhis (1997). "Linguistic Landscape and Ethnolinguistic Vitality. An Empirical Study." In: *Journal of Language and Social Psychology* 16.1, pp. 23–49.
- Latham, Kevin and Bin Wu (2013). *Chinese Immigration into the EU. New Trends, Dynamics and Implications*. London: Europe China Research and Advice Network.
- Leeman, Jennifer and Gabriella Modan (2009). "Commodified language in Chinatown. A Contextualized Approach to Linguistic Landscape." In: *Journal of Sociolinguistics* 13.3, pp. 332–362.
- (2010a). "Selling the City. Language, Ethnicity and Commodified Space." In: *Linguistic Landscape in the City*. Ed. by Elana Shohamy et al. Bristol, Buffalo, and Toronto: Multilingual Matters, pp. 182–198.
- (2010b). "Trajectories of Language. Orders of Indexical Meaning in Washington, DC's Chinatown." In: *Re-Shaping Cities. How Global Mobility Transforms Architecture and Urban Form*. Ed. by Michael Guggenheim et al. London and New York: Routledge, pp. 167–188.
- Lipovsky, Caroline and Wei Wang (2019). "Wenzhou Restaurants in Paris's Chinatowns. A Case Study of Chinese Ethnicity Within and Beyond the Linguistic Landscape." In: *Journal of Chinese Overseas* 15, pp. 202–233.
- Liu, Hong (2005). "Explaining the Dynamics and Patterns of Chinese Emigration since 1980. A Historical and Demographic Perspective." In: *Journal of Oriental Studies* 39.1, pp. 92–110.
- Lou, Jia (2007). "Revitalizing Chinatown into a Heterotopia. A Geosemiotic Analysis of Shop Signs in Washington, DC's Chinatown." In: *Space and Culture* 10.2, pp. 170–194.
- Lou, Jia Jackie (2010). "Chinese on the Side. Marginalization of Chinese in the Linguistic and Social Landscapes of Chinatown in Washington, DC." In: *Linguistic Landscape in the City*. by Elana Shohamy et al. Bristol, Buffalo, and Toronto: Multilingual Matters, pp. 96–114.
- Mao, Shangwen 毛上文 and Wen Fang 温芳 (2003). 起名技巧大全 [*Complete Collection of Naming Techniques*]. Beijing: 气象出版社 [Qixiang Chubanshe].
- Moore, Markéta and Czeslaw Tubilewicz (2001). "Chinese Migrants in the Czech Republic. Perfect Strangers." In: *Asian Survey* 41.4, pp. 611–628.
- Obuchová, Lubica (2002). *Čínská komunita v České republice 2001 [Chinese Community in the Czech Republic 2001]*. Praha: Orientální ústav AV ČR.
- Reh, Mechthild (2004). "Multilingual Writing. A Reader-Oriented Typology—with Examples from Lira Municipality (Uganda)." In: *International Journal of the Sociology of Language* 170, pp. 1–41.
- Scollon, Ron and Susie Wong Scollon (2003). *Discourses in Place. Language in the Material World*. London and New York: Routledge.

- Shang, Guowen and Libo Guo (2017). "Linguistic Landscape in Singapore. What Shop Names Reveal about Singapore's Multilingualism." In: *International Journal of Multilingualism* 14.2, pp. 183–201.
- Slaměniková, Tereza (2023). "A Touch of Chinese Culture in the Czech Public Space. Chinese Restaurant Names in Prague." In: *Onomastics in Interaction with Other Branches of Science*. Ed. by Urszula Bijak, Paweł Swoboda, and Justyna B. Walkowiak. Krakow: Jagiellonian University Press, pp. 477–500.
- Sluka, Nikolai A., Andrei V. Korobkov, and Pavel N. Ivanov (2018). "The Chinese Diaspora in the EU Countries." In: *Baltic Region* 10.3, pp. 80–95.
- Spolsky, Bernard (2009). "Prolegomena to a Sociolinguistic Theory of Public Signage." In: *Linguistic Landscape. Expanding the Scenery*. Ed. by Elana Shohamy et al. New York and London: Routledge, pp. 25–39.
- Wachendorff, Irmi (2020). "Typographetics of Urban Spaces. The Indication of Discourse Types and Genres Through Letterforms and Their Materiality in Multilingual Urban Spaces." In: *Proceedings of Grapholinguistics in the 21st Century, 2020*. Ed. by Yannis Haralambous. Vol. 4. Grapholinguistics and Its Applications. Brest: Fluxus Editions, pp. 361–415.
- Williams, C. A. S. (1976). *Outline of Chinese Symbolism and Art Motives*. New York: Dover Publications.
- Zhang, Hui, Ruanni Tupas, and Aman Norhaida (2020). "English-dominated Chinatown. A Quantitative Investigation of the Linguistic Landscape of Chinatown in Singapore." In: *Journal of Asian Pacific Communication* 30.1/2, pp. 273–289.

# Sentence-Final Particle vs. Sentence-Final Emoji

## The Syntax-Pragmatics Interface in the Era of Computer-Mediated Communication

Chenchen Song

*Abstract.* In this article, I present a formal linguistic analysis of affective emojis (i.e., emojis that are used to add tones to text messages) in computer-mediated communication (CMC) and lay out some preliminary thoughts on CMC linguistics. My analysis, which builds on the root-based approach to semilexical elements in generative syntax, separates CMC data with affective emojis into a non-CMC-specific part (i.e., the linguistic text) and a CMC-specific part (i.e., the emoji), with the latter functionally wrapping around the former and thereby setting its tone. This analysis can be applied to other CMC-specific affective elements too, such as memes and background music. The special nature of the digital modality has nontrivial ramifications for CMC linguistics. I argue that until the “legibility conditions” of the cyber-digital system are ascertained, the safest linguistic tools to use in research on CMC-specific phenomena are those that are not designed exclusively for the cognitive domain of language.


### 1. Introduction<sup>1</sup>

Haralambous (2020, p. 12) introduces grapholinguistics as “the discipline dealing with the study of the written modality of language” and

---

Thanks to the audience at the 2022 Grapholinguistics in the 21st Century Conference (6/8/2022) and the audience at the Cambridge SyntaxLab (6/28/2022) for constructive feedback. Thanks to Xiaoke Bu, Chunan Li, Shangze Li, Li Nguyen, Michele Sanguanini, Ke Wu, and Ruikang Zhang for participating in my survey.

---

Chenchen Song  0000-0002-3543-8489  
Zhejiang University  
E-mail: cjs021@zju.edu.cn

1. Abbreviations: AP = affective punctuation, C-D = cyber-digital, CL = classifier, CMC = computer-mediated communication, Conj = conjunction, CP = complementizer phrase, CRS = currently relevant state, DECL = declarative, DISP = disposal, DP = determiner phrase, EMPH = emphasis, EP = emotion phrase, MP = modal particle, NP = noun phrase, NumP = number phrase, PL = plural, POSS = possessive, Q = question marker, REL = relative clause marker, SFE = sentence-final emoji, SFP = sentence-final particle, TP = tense phrase, TU = text unit, *v*\*P = transitive light verb phrase, VP = verb phrase

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 157–192. <https://doi.org/10.36824/2022-graf-song>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

points out that the reason why it has received little recognition is because writing has long been viewed “just as an accidental secondary representation of language.” This position dates back to at least Ferdinand de Saussure’s *Course in General Linguistics* (originally published in 1916):

Language and writing are two distinct systems of signs; the second exists for the sole purpose of representing the first. The linguistic object is not both the written and the spoken forms of words; the spoken forms alone constitute the object. (Saussure, 2011)

I agree with Saussure. It is a basic fact that human language, either spoken or signed, does not depend on writing. That said, however, I wonder whether Saussure would still have put his view in such an absolute tone if he had had the chance to time-travel to the 2020s and see how human beings are staying in touch nowadays.

Face-to-face (or voice-to-voice) communication is certainly still with us, but in the meantime, modern technology has made computers, smartphones, and the like an indispensable additional channel of communication. Given this revolutionary change of lifestyle, it is unclear to me to what extent we can confidently assert that writing—or really typing (e.g., texting, tweeting)—is still strictly secondary to oral language. Among others, many CMC-specific communicative elements—such as emojis, memes, and GIFs—have never existed in oral speech and never will. They are native to the digital modality of communication instead. In this article, I present an emoji-centered case study of CMC and hope to convince readers that we need to rethink the relation between language and writing/typing in the 21st century.

Emojis play an increasingly important role in our day-to-day lives, in that they compensate for the lack of nonverbal or “paralinguistic” (Carey’s 1980 term) cues in online textual communication. As suggested by Gawne and McCulloch (2019), the place of emojis in computer-mediated communication (CMC) is equivalent to that of “tone of the voice and body language in face-to-face communication.” It is fair to say that emojis are becoming an integral part of human language in the digital age. As a linguist, I am most interested in the following questions:

1. What is the cognitive nature of CMC data involving emojis? Is the normal tool kit from linguistics sufficient for an adequate analysis of them?
2. If it turns out that the nature of CMC data is fundamentally different in certain aspects from that of conventional linguistic data, then which part of the linguistic tool kit is still applicable to their analysis?

The rationale behind these questions is as follows. Modern linguistics, in particular its generative branch (Chomsky, 1957 et seq.), is established on the hypothesis that our language capacity is supported by a

dedicated mental organ—the language faculty. This is a computational system that generates complex structures out of basic linguistic units (e.g., words). The language faculty interfaces with two other cognitive systems: the sensorimotor system and the conceptual-intentional system (Chomsky, 1995). The former is where abstract linguistic structures get externalized as physical signals, and the latter is where they get interpreted as language-based thoughts. A major goal of contemporary theoretical linguistics is to specify how information flows from the computational system to the interface systems. For instance, linguists have proposed many operations in the past few decades to tackle the question of how hierarchical syntactic structures are converted to linear strings usable in the oral-auditory modality (see Biberauer and Roberts, 2013 for an impression of the complexity of this issue). Due to the central status of linearization in pre-CMC-era linguistics, quite a few theoretical tools initially designed for linearization purposes alone have subsequently been made part of the core design of the language faculty (such as “cyclic spell-out” and its latest incarnation Phase Theory; Chomsky, 2001).

My questions above are based on the concern that, if CMC is not confined by the naturally evolved communicative modalities (including but not limited to the oral-auditory modality) or their requirements, then what theoretical linguistic tools can we still apply to CMC data, and what tools must we refrain from using? These are big questions whose settling calls for much more research and community efforts. For the limited purpose of this article, I wish to demonstrate the applicability of just one formal linguistic tool: root categorization.

As has been mentioned, my case study is centered on emojis. In particular, the emoji usage described above is *affective* in nature. Affective emojis convey speaker attitudes or tones. Emojis can also be used in a *nonaffective* way. This is the situation where an emoji is simply used as an icon for a verbal concept, usually directly substituting for a word. See (1) for an illustration.<sup>2</sup>

- (1) a. Great idea 👍 I'm in 😊  
 b. If I were in Detroit, I'd give you a 🎁 (adapted from Maier, 2021, p. 4)

The two emojis in (1a) are used affectively. They respectively express an approving tone and a genuinely happy tone. By contrast, the emoji in (1b) is used nonaffectively. It merely represents a gift and can be directly replaced by the word “gift.” The two types of emoji usage above may be alternatively described as use-conventional vs. truth-conditional

---

2. I generally use Apple emojis in this article but will switch to alternative versions in cited examples, since different implementations of the same emoji often have subtle differences in the exact affects they convey (see §2.3).

or non-at-issue vs. at-issue (Grosz, Greenberg, De Leon, and Kaiser, 2021, Maier, 2021, Pierini, 2021), the latter being based on a piece of terminology in Potts (2005). In what follows, I will stick to the affective vs. nonaffective terminology.

I focus on affective emojis in this article. Note that the two affective emojis in (1a) are both attached to the end of the sentence they accompany—or more exactly the *text unit*, since “Great idea” is not a complete sentence. This syntactic property is true of affective emojis in general. Hence, I also call affective emojis *sentence-final emojis* (SFEs). I choose this designation because the above combination of syntactic and semantic properties—namely, being sentence-final and expressing speaker affects—is reminiscent of a class of vocabulary elements in oral languages, especially in East and Southeast Asian languages, which have been called “sentence-final particles” (SFPs) in the linguistic literature (see, e.g., Cheng and Tang, 2022 and Morita, 2018). See (2) for some examples from Mandarin Chinese, which is also my main source of data.<sup>3</sup>

- (2) a. *xià xuě le ye* [Mandarin Chinese]  
 fall snow CRS SFP  
 ‘It snowed. (excited tone)’
- b. *xià xuě le a*  
 fall snow CRS SFP  
 ‘It snowed. (surprised tone)’
- c. *xià xuě le you*  
 fall snow CRS SFP  
 ‘It snowed. (kind reminder tone)’
- d. *xià xuě le ha*  
 fall snow CRS SFP  
 ‘It snowed. (harmony-seeking tone)’

In (2), the same new situation “it snowed” is reported in four different tones, which are encoded in four different SFPs. In CMC, the same communicative effects can be achieved via affective emojis, as in (3).

- (3) *xià xuě le* 😄/😲/😏/😌 [Mandarin Chinese]  
 fall snow CRS SFE  
 ‘It snowed. (excited/surprised/reminder/harmony-seeking tone)’

The particle-emoji parallelism above is striking. One may even conclude that SFEs are the digital counterpart of SFPs. Indeed, the two types of

---

3. I follow the standard practice in linguistics and present non-English examples in a three-line format: the first line is the original example (or its romanization, if the original language has a non-Latin script), the second line is a verbatim glossing of the example (in an English-based metalanguage), and the third line is a more natural English translation.



affect-expressing elements have been given a unified linguistic analysis in Song (2019). However, in this article I will show that despite their functional similarity, we cannot put SFPs and SFEs in the same category. While the former are an integral part of oral speech, the latter are first-class citizens of CMC (and CMC alone). I will present three arguments that bear out the categorial distinction between SFPs and SFEs:

1. SFPs and SFEs can and often do co-occur.
2. SFPs are a closed class, whereas SFEs are an open class.
3. The positioning of affective emojis is not influenced by crosslinguistic word order variation, whereas that of affective particles is.

The three arguments will be elaborated one by one. After that, I will propose a new linguistic analysis for SFEs, which is based on the Generalized Root Syntax theory in Song (*ibid.*).

The rest of this article is structured as follows. In Section 2, I present my arguments against an identical linguistic treatment of SFPs and SFEs. In Section 3, I present my new analysis of affective emojis. In Section 4, I discuss the implication of my case study for the field of CMC linguistics in general. Section 5 concludes.

## 2. SFP and SFE are different categories

In this section, I comparatively examine the linguistic behavior of SFPs and that of SFEs and argue that they should not be treated as the same category. I begin with a note on SFP taxonomy (§2.1), then move on to present my three arguments (§2.2–2.4), and finally make a digression on sentence-initial emojis (§2.5), showing that they are not counterexamples to my generalization. I end the section with an interim summary (§2.6) that prepares the ground for my theoretical analysis.

### 2.1. SFP taxonomy

SFPs are not a homogeneous category. According to Paul (2014), the SFPs in Mandarin Chinese fall in three types, as shown in Table 1.

Type I SFPs in Mandarin are tense or aspect markers, such as the currently relevant state marker *le*, the effect of which partly overlaps with that of the perfect in English. Thus, in the “snowing” examples in (2), a more accurate paraphrase of the statement “it snowed” is “it has snowed some time ago, and that state of affairs is relevant to the current situation we are in (e.g., there is snow on the ground).” Type II SFPs are sentence type markers, such as the yes-no question marker *ma*, which turns a proposition into a yes-no question and is similar in effect to French *est-ce que*. Thus, while *xià xuě le* ‘it snowed’ is a statement, *xià*

TABLE 1. A taxonomy of Mandarin Chinese SFPs (adapted from Paul, 2014)

| Type | Characterization | Examples   |
|------|------------------|--|
| I    | Tense/Aspect     | <i>le</i> ‘currently relevant state’<br><i>lázhibe</i> ‘recent past’<br><i>ne</i> <sub>1</sub> ‘continued state’ |
| II   | Sentence type    | <i>ma</i> ‘yes-no question’<br><i>ba</i> ‘imperative’<br><i>ne</i> <sub>2</sub> ‘follow-up question’             |
| III  | Attitude         | <i>o</i> ‘mild reminder tone’<br><i>a/ya</i> ‘surprised tone’<br><i>ne</i> <sub>3</sub> ‘exaggerating tone’      |

*xuě le ma* ‘It snowed?’ is a question. Type III SFPs are attitude markers. All four examples in (2) are of this type. This is also the type of SFP that I focus on in this article. Hereafter, by “sentence-final particle” I only mean Type III SFPs.

## 2.2. Argument I: SFPs and SFEs can co-occur

The first reason why SFPs and SFEs should not be treated as the same category is that they can and often do co-occur in the same sentence. For instance, the patterns in (2) and (3) can be combined into (4).

- (4) a. *xià xuě le ye* 😄 [Mandarin Chinese]  
 fall snow CRS SFP SFE  
 ‘It snowed. (excited tone)’
- b. *xià xuě le a* 😲  
 fall snow CRS SFP SFE  
 ‘It snowed. (surprised tone)’
- c. *xià xuě le you* 😊  
 fall snow CRS SFP SFE  
 ‘It snowed. (kind reminder tone)’
- d. *xià xuě le ha* 🙏  
 fall snow CRS SFP SFE  
 ‘It snowed. (harmony-seeking tone)’

In fact, the forms in (4) are more natural than those in (3), because the retention of the SFPs makes the messages more speech-like, while the addition of the SFEs helps further highlight the tones in the SFPs. Such SFP-SFE co-occurrence is common in CMC data. See (5) for more examples from the social media website Sina Weibo (henceforth Weibo), which is the Chinese equivalent of Twitter.

- (5) a. *wǒ měitiān dōu zài zhíbō o qīn 😊* [Mandarin Chinese]  
 I everyday all be.at live-stream SFP dear SFE  
 ‘For your information, dear, I’m live-streaming everyday. (teasing tone)’
- b. *nǐ de wǎng-míng běn fúhé nǐ o 😊*  
 you POSS Internet-name very suit you SFP SFE  
 ‘Just saying, your profile name suits you very well. (jocularly cheeky tone)’
- c. *wǒ zěnmē jìde bǎoxiàng shì liú bǎ tā chuài le a 🤔🤔🤔*  
 I how remember likely is Liu DISP her dump CRS SFP SFE  
 ‘How come I vaguely remember that it was Liu who had dumped her? (highly amused tone)’
- d. *nǚ míngxīng shēngrì kuàilè o 🥰*  
 female star birthday happy SFP SFE  
 ‘Superstar girl, happy birthday! (cute fangirl tone)’ (Weibo)

Like many Asian social media platforms, Weibo has its own emojis, which are outside the Unicode list. Nevertheless, the usage of the Weibo-specific emojis in (5) is not different from that of the Unicode emojis we have seen. Moreover, in these examples, the SFEs are not translations of the SFPs. Rather, in each example, the affects in the SFP and the SFE combine into a new and more subtle tone. I will come back to platform-specific, non-Unicode emojis in Section 2.3. Specifically, (5a), (5b), and (5d) share the same base tone—the mild reminder tone encoded in the SFP *o*—which is further shaped by the additional SFEs in three different ways, respectively into a teasing reminder, a jocularly cheeky reminder, and a fangirlish reminder. Similarly, the surprised-tone SFP *a* in (5c) combines with the “allow me to do a sad face” emoji (repeated three times) to yield a seemingly surprised but actually highly amused tone.

The productive co-occurrence of SFPs and SFEs is a clear indication that the two types of affective element instantiate different linguistic categories, with a category being understood as an equivalence class in terms of linguistic behavior. To begin with, linguistic elements of the same category are usually in complementary distribution, which is partly what motivates linguists to define them as a category in the first place. See (6) for two familiar examples. The asterisk indicates that the expression after it is ill-formed.

- (6) a. this book, that book, \*this that book (Demonstrative)  
 b. I like reading, you like reading, \*I you like reading (Pronoun)

*This* and *that* are in the same category (Demonstrative) because they can freely substitute for each other without affecting grammaticality and cannot be used simultaneously, and the same is true for the nominative pronouns *I* and *you*. Note that the conception of category adopted here is

a fine-grained one. Assuming categories are hierarchically organized in their ontology into super- and subcategories, I only consider elements of the same smallest subcategory as categorially equivalent. Thus, while nonnominative pronouns like *me* and *him* are also in the general category Pronoun, they are not equivalent to nominative pronouns.

Furthermore, when SFPs and SFEs co-occur, their order cannot be switched. That is, the SFP slot can only be filled by oral-language particles, while the SFE slot can only be filled by emojis (or other similar digital symbols, such as emoticons). Sentences like the following are unacceptable.

- (7) a. \**xià xuě le 😊 ye* [Mandarin Chinese]  
       fall snow CRS SFE SFP  
       ‘It snowed. (excited tone)’
- b. \**nǐ de wǎng-míng hěn fúhé nǐ 😊 o*  
       you POSS Internet-name very suit you SFE SFP  
       ‘Just saying, your profile name suits you very well. (jocularly cheeky tone)’

This restriction is unexpected if the two types of affective elements are categorially equivalent.

### 2.3. Argument II: SFEs are an open class

The second reason why SFPs and SFEs should not be treated as the same category is that SFPs are a closed class, while SFEs are an open class. Thus, even if they were in the same category, that category would still be a hybrid one, with two heterogeneous subcategories, which brings us back to the ontological issue mentioned above.

The inventory of SFPs in Sinitic languages is not particularly small, especially if we take all three subtypes in Section 2.1 into consideration. However, they are still a closed class, which means that the set of SFPs in a Sinitic language is stably fixed in an extended period of time. Take Mandarin Chinese for example. Although scholars hold varied opinions on the number of SFPs it has, that number is generally assumed to be under 30. Among others, Chao (1968) lists 26 (including many borderline items), Sun (1999) lists 28 (for all Mandarin subvarieties throughout the 19th and 20th centuries), and Li and Thompson (1981) list 6 (only the most common ones).

By contrast, the inventory of SFEs is much larger and also keeps expanding. This is evidenced by four observations:

1. New face emojis are created every year.
2. Nonface emojis can be used affectively too.
3. There are plenty of platform-specific, non-Unicode affective emojis.

4. There are various quasi emojis (e.g., emoticons, affective punctuation marks).

In what follows, I will elaborate on these observations one by one. First, new face emojis are being regularly created, almost on a yearly basis. See (8) for some examples.

- (8) 2018: 😊, 😊, 😊, 😊, 😊, 😊  
 2019: 😊  
 2020: 😊, 😊, 😊, 😊, 😊, 😊  
 2021: 😊, 😊, 😊, 😊, 😊, 😊, 😊, 😊 (Emojipedia)

Face emojis are naturally affective, so their constant expansion is clear evidence of the open-class nature of SFEs. However, not all affective emojis are face emojis, and that brings us to the second piece of evidence listed above—namely, that nonface emojis can also be used affectively. When studying affective emojis, we should not limit our attention to just face-based ones (pace Grosz, Greenberg, De Leon, and Kaiser, 2021).

The affective use of nonface emojis is highly versatile. Some more systematic ones are hand emojis like 🙌, 🙏, and 🙌 and heart emojis like ❤️, ❤️, and ❤️. There are also less systematic ones, such as those in (9). For authenticity's sake, I have retained the spelling and emoji style (i.e., the Twitter version) of the original tweets.

- (9) a. Perfect art! So talented artist 🔥  
 b. had 'hug' been a little more second longer, she would've elbowed one of these queens out. just saying 💅  
 c. Every woman wants a man who's hard-working and ambitious until it's the weekend and he plans on working 🐸🍷 (Twitter)

In all these examples, the nonface emojis are clearly used affectively, in that they serve to convey speaker attitudes. The fire emoji in (9a) conveys an enthusiastically admiring tone, the nail polish emoji in (9b) conveys a nonchalant tone, and the frog-and-hot-beverage emoji compound in (9c) conveys a sarcastic tone.<sup>4</sup> While the above affective uses are all largely conventionalized—in that they are regularly used in the relevant affective senses—there are also more ad hoc affective uses of nonface emojis. The temporary nature of such usage is especially clear in cases where multiple emojis are randomly put together to convey a strong emotion, as exemplified in (10).

4. While the tones in 🔥 and 💅 are stably fixed, the tone in 🐸🍷 is less so. According to Emojipedia, this combination could be used for gossip or sarcasm or be associated with trolling or the alt-right.



- (13) a. *xiànzài bàochū shénme xīnwén wǒ dōu bù xīqí le* 🙄  
 now break whatever news I EMPH not curious CRS SFE  
 ‘Nowadays I’m no longer shocked by whatever news. (onlooker’s tone)’
- b. *wúcháng fēnxiāng gěi nimen* 🙄  
 gratis share to you all SFE  
 ‘I’m sharing these (celebrity scandals) with you for free. (onlooker’s tone)’ (Weibo)

Note that the three platform-specific implementations of the watermelon-eating emoji mentioned above are not completely equivalent. Intuitively, the Weibo version 🙄 has a more “none of my business” attitude, the WeChat version 🍉 has a more gossipy feeling, while the Douyin version 🍉 has a more “peanut gallery” effect. Such subtle tonal variation in a sense makes the inventory of SFEs even larger—because if two variants of the same emoji convey different tones, then they may as well be treated as different emojis.

Internet users’ intuition over the tonal variation in affective emojis is impressively nuanced. I conducted a small-scale survey on whether the platform-specific implementations of the eye-rolling emoji convey the same tone, and the general answer I got was No. See Table 2 for the detailed responses. Note that the two QQ<sup>5</sup> versions are both animated, but I can only present them as static screenshots here. To further illustrate the rich intuition Internet users possess over emoji usage, I quote the following additional comments from my respondents:

Compared with the other eye-rolling emojis, this animated one [QQ 2] ... adds extra absurdity and humor. With the smiling, there is also a slightly sarcastic tone. I think it is a mixture of complex emotions and subtle feelings. Thus, personally, I find it peculiarly lovely. (User 5)

For me, emojis with a nonflat mouth are more negative than emojis with a flat mouth, which are in turn more negative than emojis with an open mouth. So, here the Twitter version of the eye-rolling is more negative than the Apple version, which in turn is more negative than the first WeChat version. The second QQ version is different from all the others. I tend to express the emotion of sarcasm or fake politeness when using it. (User 7)<sup>6</sup>








Finally, the abundance of affective quasi emojis, such as emoticons and affective punctuation marks, is also evidence that SFEs are an open class. Modern-day emoticons are far more versatile than sideways smileys like :- ) and :D. Once again, Asian Internet users are particularly creative in this realm. See (14) for some examples of Japanese kaomojis.<sup>7</sup>

5. QQ is a Chinese instant messaging software service.

6. Since User 7 only provided this general remark, I did not include their response in Table 2.

7. These examples are extracted from [kaomoji.ru/en/](http://kaomoji.ru/en/). (last visited on 10/22/2022)

TABLE 2. Survey results on the tonal differences between platform-specific eye-rolling emojis

|        |  | Apple |  | Twitter              |  | WeChat 1                     |  | WeChat 2               |  | Weibo                           |  | QQ1                                  |  | QQ2                                   |
|--------|---|-------|---|----------------------|---|------------------------------|---|------------------------|---|---------------------------------|---|--------------------------------------|---|---------------------------------------|
| User 1 | "I can't even," jaded   |       |   | disappointed         |   | "eye-avoidance," embarrassed |   | disappointed           |   | disappointed & sad              |   | slightly embarrassed or a bit cheeky |   | amused (for chaos or minor confusion) |
| User 2 | slightly noyed  | an-   |   | a bit sad            |   | wondering                    |   | confused               |   | slightly different or skeptical |   | in-slightly naughty                  |   | silly                                 |
| User 3 | speechless (negative)   |       |   | negative attitude    |   | playing innocent, "not me"   |   | pretending to be angry |   | negative attitude               |   | playing innocent, "not me"           |   | speechless (negative)                 |
| User 4 | speechless  |       |   | speechless & unhappy |   | "I don't wanna hear"         |   | pretending to be angry |   | speechless (friendlier)         |   | "I don't wanna hear" (cuter)         |   | totally speechless, "death smile"     |
| User 5 | real rolling (highly negative)  | eye-  |   | ≈Weibo               |   | ≈QQ1                         |   | pretending to be angry |   | a bit of disdain                |   | a bit shocked                        |   | humorously sarcastic                  |
| User 6 | real rolling  | eye-  |   | confused             |   | pretending to be confused    |   | arrogant               |   | pondering                       |   | pretending to be confused            |   | backhanded compliment                 |



|      |                 |             |                 |               |            |
|------|-----------------|-------------|-----------------|---------------|------------|
| (14) | (~^)            | (*^▽^~)b    | (.♣.♣.)         | (// ^ ▽ ^ //) | o(*.♣.)o   |
|      | (o~^o)          | o(≡▽≡)o     | ('♣.~.♣.')      | \(★ω★)/       | \(≡▽≡)/    |
|      | \(o^ ^ ▽ ^ o)/  | \(o^ ^ o)/  | (*^~^*)         | \(▽^▽^)       | (≡~≡)      |
|      | ((o(*^ ▽ ^ o))) | (o'▽'o)     | (@ ^ ^ ^)       | (o.ω.o)       | (~ω~)      |
|      | ('♣.ω.♣.')      | \( . ▽ . )/ | ♣.::★*.~(*^~^*) | <(▽^▽^)>      | (o^ ▽ ^ o) |

Kaomojis are highly popular in Japan and China—so much so that smart-phone keyboards now have a special section for them. In addition, Chinese and Korean speakers sometimes use special emoticons made up of Chinese/Korean characters, such as 囧 (an embarrassed face) in Chinese<sup>8</sup> and ○ □ ○ (a shocked face) in Korean.

As for affective punctuation marks, apart from the conventional question and exclamation marks (and their various combinations), the ellipsis and the tilde are good examples too. The former is popular around the world and usually signals hesitation or silence, while the latter is mainly popular in Asia and signals cuteness or a softened tone. See (15) for an illustration.<sup>9</sup>

- (15) a. *zhēnde ma ...* [Mandarin Chinese]  
           real Q AP  
           'Really? (hesitant tone)'
- b. *bāng wǒ mǎi dōngxì ~~~*  
           help me buy stuff AP  
           'Help me buy something please. (cute tone)'

The ellipsis in (15a) conveys hesitation, which may be translated as “alright” or “whatever” depending on the context. The tildes in (15b), on the other hand, create a friendly and cute-sounding effect, which is important in texting since otherwise the message sounds rather abrupt.

#### 2.4. Argument III: Affective emoji positioning is not influenced by word-order variation

My third argument for the categorial difference between SFPs and SFEs is based on a more general observation about affective particles. SFPs in Chinese and other Asian languages are a major type of affective particle in human language, but they are not the only type. Among others, the

8. The Chinese character 囧 originally means “bright” and is pronounced *jiǒng*, but its usage as an emoticon has nothing to do with its original meaning and is merely a shape-based recycling.

9. Chinese speakers often use a sequence of Chinese-style periods (。。。) in place of ellipsis dots (...), and for some speakers the former conveys an even stronger hesitant tone. I abstract away from this subtlety here.

TABLE 3. Some German modal particles (Durrell, 2021, §9.1)

| Particle    | Connotation   |
|-------------|---|
| <i>balt</i> | an attempt by the speaker to put an end to any discussion because the situation does not allow any alternatives                   |
| <i>ja</i>   | appealing for agreement, expressing surprise, intensifying commands   |
| <i>mal</i>  | making the tone sound less blunt  |
| <i>doch</i> | typically used to try to persuade the listener of the speaker's point of view, usually expressing a contradiction or disagreement |
| <i>nun</i>  | signaling dissatisfaction with a previous answer or that the speaker considers the topic exhausted                                |
| <i>eben</i> | typically expressing a confirmation that something is the case  |

modal particles in German (and some other Germanic languages, such as Dutch; see, e.g., Fehringer and Cornips, 2019) are also affective, in that they also serve to convey speaker tones or attitudes. See Table 3 for a selection of common German modal particles and see (16) for some concrete examples.

- (16) a. *Das ist **balt** so.* [German]  
that is MP so  
‘But there, that’s how it is. (there’s-nothing-one-can-do tone)’  
b. *Ihr **habt ja** früher zwei Autos gehabt.*  
you.PL have MP earlier two cars had  
‘Of course, you used to have two cars. (as-we-all-know tone)’  
c. *Ich kann ihn nicht überreden. Er ist **eben** hartnäckig.*  
I can him not convince he is MP obstinate  
‘I can’t convince him. He’s just obstinate. (it-can’t-be-helped tone)’  
(Durrell, 2021, §9.1)

As we can see, the position of affective modal particles in German is consistently sentence-medial instead of sentence-final. This shows that the syntactic position of affective particles, like that of most other elements of oral languages, is subject to crosslinguistic variation. By contrast, the positioning of affective emojis does not follow this general observation. They are regularly sentence-final even in German, as in (17).

- (17) a. *Ich wünsche euch einen guten Morgen!* 😊😊😊 [German]  
I wish you.PL a good morning SFE  
‘I wish you all a good morning! (very friendly and blissful tone)’  
b. *Ich würde mehr Geld als in meinem Vollzeitjob machen* 🤖  
I would more money than in my full time job make SFE  
‘I’d make more money than in my full time job. (shocked tone)’  
(Twitter)

Some speakers even view modal particles as “verbal emojis,” as reflected in the two online remarks below:

Modal particles are little words that express connotations such as feelings or moods. Because of this, they are also sometimes referred to as “filler words.” Basically, they amount to verbal emojis :D (chatterbug.com<sup>10</sup>)

IMO the most important thing to understand about modal particles is that they change mood, not meaning. They are effectively “verbal emojis.” (soupsticle on Reddit<sup>11</sup>)

The Reddit user in the second quote above further illustrates their point with the similarity between the modal particle *halt* and the shrug emoji 🤷, as in (18).

- (18) a. *Das ist halt so.* = That's how it is. 🤷  
 b. *Dann hat er halt eine große Nase.* = So he has a big nose, so what? 🤷

The syntactic heterogeneity of affective modal particles and affective emojis in German is most clearly seen when they co-occur in the same sentence, as in (19).

- (19) a. *Nachts ist ja eine Menge los, dafür muss er ja* [German]  
 at night is MP a lot going on therefore must he MP  
*tagsüber sehr viel schlafen* 😊  
 during the day very much sleep SFE  
 ‘There’s a lot going on at night, so he (the speaker’s cat) has to sleep a lot during the day. (humorously as-we-all-know tone)’  
 b. *Wieso ist dir das denn so wichtig?* 🙄  
 why is to you that MP so important SFE  
 ‘Why is that so important to you? (nonchalantly obliging tone)’  
 (Twitter)

As we can see, the affects in the modal particles and the SFEs add up, just like the situation in Chinese sentences with both SFPs and SFEs (see (5)). In (19a), the modal particle *ja* (which occurs twice) conveys an agreement-seeking, as-we-all-know tone, and the SFE 😊 further adds some minor awkwardness and embarrassment to it (because the speaker’s cat sleeps all day long), thus making the overall tone of the tweet humorously fake-serious. Likewise, in (20b) the modal particle *denn* serves to make the question more obliging (and less blunt), while

10. <https://chatterbug.com/grammar/german/modal-particles-modalpartikeln> (last visited on 10/22/2022)

11. <https://www.reddit.com/r/German/comments/qmit3d/comment/hj9t3f1/> (last visited on 10/22/2022)

TABLE 4. A crosslinguistic survey of affective emoji positioning

| Language  | Family           | Type          | Basic w.o. Aff. emoji position |
|-----------|------------------|---------------|--------------------------------|
| Mandarin  | Sinitic          | isolating     | SVO sentence-final             |
| Japanese  | Japonic          | agglutinative | SOV sentence-final             |
| Korean    | Koreanic         | agglutinative | SOV sentence-final             |
| English   | Germanic         | analytic      | SVO sentence-final             |
| German    | Germanic         | fusional      | SOV sentence-final             |
| French    | Romance          | fusional      | SVO sentence-final             |
| Irish     | Celtic           | fusional      | VSO sentence-final             |
| Basque    | language isolate | agglutinative | SOV sentence-final             |
| Hungarian | Finno-Ugric      | agglutinative | rel. free sentence-final       |

TABLE 5. Illustration of affective emoji positioning across languages

| Language  | Example   |
|-----------|---|
| Japanese  | <i>gozenchū no ame wa dokoni ittandesu ka</i> 🤔<br>'Where did the rain in the morning go? (pondering tone)'                           |
| Korean    | <i>membeo-deul-i 'bat-gyu'-rago bureum</i> 😍<br>'The members calling him "hot-gyu." (excited fangirl tone)'                           |
| French    | <i>C'est réducteur au possible ces fêtes</i> 😞<br>'These holidays are as simplistic as possible (frustrated tone)'                    |
| Irish     | <i>RT agus fág trácht le bbeith san áireamb!!</i> 🥰<br>'RT and leave a comment to be included!! (enthusiastic tone)'                  |
| Basque    | <i>Bilera eta ekitaldi nagusiak bueltan dira Euskaldunan</i> 😊<br>'Meetings and big events are back in Basque. (happy and cute tone)' |
| Hungarian | <i>Sajnos nem tehetek többet</i> 🙔<br>'Unfortunately I can't do more. (sad tone)'   |

the SFE 🙔 adds a nonchalant coloring to the interrogation, thus making the overall tone of the tweet humorously aloof.

To further investigate the syntactic position of affective emojis across languages, I examined posts in nine languages on Twitter and Weibo, as summarized in Table 4. The results show that regardless of the variation in language type and basic word order, affective emojis are invariably sentence-final. See Table 5 for a crosslinguistic illustration (except English, Chinese, and German, which we have seen examples of).

The insensitivity of affective emoji positioning to crosslinguistic word order variation is even more evident in cases where the same content is posted in two languages, as in the Basque and Spanish tweets in (20).

- (20) a. *Bilera eta ekitaldi nagusiak bueltan dira Euskaldunan* 😊 [Basque]  
*Los grandes eventos y las reuniones están de vuelta* [Spanish]  
*en Euskalduna* 🎉  
 ‘Meetings and big events are back in Basque. (happy and cute tone)’
- b. *Bizkaia egunero zaintzen ditu mendetasun-egoeran dauden* [Basque]  
*adineko milaka pertsona* 🧓🧓❤️  
*Bizkaia cuida cada día de miles de personas mayores en situación* [Spanish]  
*de dependencia* 🧓🧓❤️  
 ‘Every day, Bizkaia cares for thousands of elderly people in a situation of dependency. (senior-citizen-loving tone)’ (Twitter)

In sum, since SFPs and SFEs have clear distinctions in their syntactic behavior, we cannot treat them as elements of the same category.

## 2.5. Sentence-initial emojis

In the foregoing discussion, I have made the generalization that affective emojis are consistently sentence-final across languages. However, there are also sentence-initial emojis that to some extent encode speaker affects. I discuss three such scenarios in this section and show that none of them is a real counterexample, as they are all qualitatively different from the kind of affective emojis we are concerned with.

### 2.5.1. Responses to earlier messages

The first type of sentence-initial affective emoji involves normal affective emojis. However, a closer examination reveals that these emojis do not really form a discourse unit with the subsequent sentence but are responses to earlier messages instead. See (21) for an illustration.













- (21) a. A: How is she 10 years older than him? She looks 10 years younger 😊.  
 B: 😂😂 From which angle does she look younger than him?  
 (YouTube)
- b. bts.bighitofficial: Left and Right (feat. Jung Kook of BTS) Release  
 — 🥰🥰🥰 another number one another national anthem 💜🌟  
 (Instagram)

In (21a), B’s use of the face-with-tears-of-joy emoji (twice) is an immediate response to A’s comment, which B finds hilarious. This usage of affective emojis is reminiscent of interjections, so the double face-with-tears-of-joy emoji can be replaced by words like “hahaha” and “LMAO,” and the response would still be felicitous if we remove the subsequent question (“From which angle...”). Similarly, (21b) is an Instagram post on the Korean boy band BTS’s account together with a fan’s comment.

The comment begins with an enthusiastic emoji response (three smiling faces with heart-eyes in a row) to the original post. Interestingly, the fan's further comment following the initial response is itself accompanied by a compound sentence-final emoji, which conveys a BTS-loving tone (the purple heart emoji is reserved for BTS in Korean pop culture). The scope difference between the sentence-initial and sentence-final emojis in (21) is intuitively clear and expected. In both cases, some verbal content is added to the discourse first, and some affective content next, with the latter being a response to or a modification of the former.




### 2.5.2. *Creative bullet list icons*

Some sentence-initial emojis are bullet list icons. They may be merely for creative visual purposes, as in (22a), or furthermore encode certain speaker attitudes, as in (22b).

- (22) a. Ronaldo at the 2002 World Cup:
-  7 appearances
  -  33.0 touches p/g
  -  8 goals
  -  34 shots/19 on target
  -  23.5%
  -  69.0 minutes per goal
  -  13 key passes
  -  7.90 average Sofascore rating
- b.  someday when i comeback to korea, i shall upload many sound-clouds
-  but is there a spare time between comeback preparation and concert
-  ah
-  comeback cancel (Twitter)

In (22a), miscellaneous emojis are used to introduce bullet points as well as to highlight their themes. These emojis are nonaffective. In (22b), the sun emoji is used by a fan to list some text messages from their idol, and this time the fancy bullet list icon is not only creative but also affective, conveying a warm and affectionate tone.

There are also bullet list icon emojis that are neither theme-specifying nor affective but deictic in nature, in that they directly point to the items they introduce, either literally or figuratively. See (23) for some examples.

- (23) a.  Special Shows
-  Food and blood donations
-  Upcoming movie posters/videos

- 🚩 Banners, bike rallies and other celebrations  
 A day with many surprises and celebrations 🎆🔥❤️...
- b. 🇹🇩 The conflict has halted aid deliveries to Tigray...  
 🇧🇪 #StopWarOnTigray  
 🇧🇪 #EritreaOutOfTigray (Twitter)

Both the point-right emojis in (23a) and the loudspeaker/speaking-head emojis in (23b) are used deictically, drawing readers' attention to the messages they introduce. Also note that the user in (23a) switches to the sentence-final position again when they intend to wrap a text unit in a certain tone—with an ad hoc emoji sequence plus an affective punctuation mark (the ellipsis).

Overall, bullet list icon emojis, whether affective or not, are qualitatively different from the text-accompanying affective emojis we are concerned with, the *major* function of which is tone-setting. As an aside, bullet list icon emojis, being consistently sentence-initial, are not subject to the kind of crosslinguistic word order variation we have observed in affective modal particles either. Their categorial status is beyond the scope of this article but should be part of a general study on emojis.

### 2.5.3. *Decorative frames*

Sometimes emojis are used for purely decorative purposes, where they provide a fancy frame for the messages or posts they accompany and thereby highlight them. See (24) for an illustration.

- (24) a. 🌟🌸🌸 Peaceful Morning 🌸🌟  
 🌸🌿❤️ Nature Peace ❤️🌿🌸
- b. 🌿❤️ DAY 1 ❤️🌿
- c. 🌟 manifesting these two for tomorrow's final 🌟
- d. 🚨 EMERGENCY 🚨  
 !! SWIFTIES LETS VOTE WHILE WAITING FOR SPOTIFY NUMBERS, WE ARE LOSING BADLY !! (Twitter)

Being part of a frame, the sentence-initial emojis in (24) are not really sentence-initial but more exactly sentence-surrounding—and they certainly are not only applicable to sentence-level text units either but can enclose any content that the speaker intends to highlight. Thus, they are like fancier versions of the more conventional emphasis asterisks often seen in e-mails, as in (25a). The two types of emphasis markers can also be used together, as in (25b).

- (25) a. I know \*nothing\* about my Indigenous roots.  
 b. 🌸 \*\*GOOD NEWS ALERT\*\* 🌸 (Twitter)

Compared to the emoji emphasis markers in (24), the asterisks in (25) are less expressive, but the two types of punctuation elements essentially work in the same way. Just like the bullet list icon emojis in Section 2.5.2, these frame emojis are not subject to crosslinguistic variation in positioning either. Beyond their primary function of emphasis scope demarcation, they may additionally encode speaker attitudes. Thus, the emoji frames in (24) respectively convey a peaceful tone, a nature-loving tone, a good-vibe tone, and an attention-craving tone. However, tone-setting is merely a *secondary* effect of frame emojis, which is again like the situation with bullet list icon emojis but not like the situation with the text-accompanying affective emojis we are concerned with in this article, the *primary* purpose of which is to set the tone for the text unit they accompany. To avoid ambiguity, we can call the latter *purely affective emojis*.

## 2.6. Interim summary

Affective emojis in CMC are similar in function to affective particles in verbal speech, such as final particles in Chinese and modal particles in German/Dutch. Despite their functional similarity, however, we cannot treat them as the same category in an adequate linguistic analysis of CMC data. First, the two types of affective element often co-occur. And when they do so, they must assume a strict order (SFP < SFE). Second, they differ in the open/closed nature of their inventory class, with SFPs being a closed class, and SFEs, an open class. In linguistic terms, this suggests that SFPs are more like a grammatical category (for function words), whereas SFEs are more like a lexical category (for content words). Third, purely affective emojis are consistently sentence-final in languages of different families and types, while the positioning of affective particles covaries with the general word order variation across languages. This suggests that SFEs and SFPs are subject to different syntactic rules, which in turn is a clear indication of their distinct categorial status. There are also sentence-initial affective emojis, but those we have observed are either responses to earlier message or have other primary functions (e.g., bullet list-creating, emphasis scope-demarcating) and hence constitute separate uses of emojis. My generalization and theorization in this article are only about purely affective emojis, whose main purpose is to give the text they accompany a certain tone.

The above properties of SFEs present a curious case for linguistic theory. On the one hand, SFEs are functionally similar to SFPs and usually accompany entire text units, which means that their place in the



syntactic structure of utterances is in the grammatical rather than the lexical domain. It is a basic assumption of modern syntactic theory that the grammatical zone of human language builds on top of the lexical zone. On the other hand, however, the open-class nature of SFEs make them more akin to a lexical category. Further evidence of their lexical status is the frequent conventionalization of their affective senses. For instance, the use of 🙄 to convey a nonchalant tone is not predictable from the face value of the emoji, nor is the use of 😏 by Chinese speakers to convey an onlooker's attitude. Such meaning conventionalization is highly similar to that in content words or idioms. For instance, that *dog* means "a type of four-legged animal" and that *let the cat out of the bag* means "to reveal a secret" are not predictable either and must be learned.

The conclusion we can draw from the foregoing discussion is that SFEs are a *semi-functional-semi-lexical* (henceforth *semilexical*) category. Hence, an adequate linguistic analysis of them should be based on a theory of such categories in general. In the next section, I will introduce such a theory.

### 3. A formal syntactic theory

Formal syntax is a branch of modern linguistics that approaches the grammatical structure of human language in a formally explicit way. Its origin (in the 1950s) was closely related to formal language theory in computer science (see, e.g., Chomsky, 1959), though nowadays most formal syntacticians have shifted the focus of their research to empirically grounded linguistic analysis. My analysis in this section is developed within the Minimalist Program (Chomsky, 1995 et seq.). I begin with a general introduction of semilexical elements in human language (§3.1) and then go on to introduce the Generalized Root Syntax theory (§3.2), which is a particular theoretical tool within the Minimalist Program. Finally, I demonstrate how this tool can help us explain the behavior of SFEs (§3.3).

#### 3.1. Semilexicality

Semilexical elements are linguistic elements (mostly words, but also affixes) with both substantive content and grammatical function. By "substantive content," I mean idiosyncratic descriptive content of various sorts. The most familiar classes of words with such content are the major parts of speech (aka lexical categories): Noun, Verb, and Adjective. For instance, *dog*, *cat*, and *bird* are all nouns and can freely substitute for one another in sentences without affecting syntactic well-formedness; they are only considered different words by virtue of their different

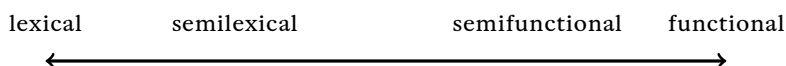


FIGURE 1. Continuum of lexicality in linguistic elements

idiosyncratic content (i.e., they name three different animals). By contrast, purely functional elements, such as the definite article *the* and the infinitive marker *to*, have no such substantive content; they only serve grammatical purposes. Words of the major parts of speech are quintessentially idiosyncratic in meaning, but idiosyncrasy can exist in other word classes too. In other words, purely lexical/functional elements are just two extremes of a continuum, as in Figure 1.

Near the lexical end of the continuum are largely lexical elements that simultaneously perform some grammatical function, such as English light verbs (26). On the other hand, near the functional end of the continuum are largely functional elements that simultaneously show some lexical idiosyncrasy, such as Mandarin Chinese conjunctions (27), the usage of which is conditioned by pragmatic factors like the formality of the context.

- (26) a. **take** a break, **make** a deal, **do** exercises [English]  
 b. *bé* ‘and (neutral)’, *gēn* ‘and (colloquial)’ [Mandarin Chinese]  
*yǔ* ‘and (formal/literary)’, *jì* ‘and (solemn)’

All three boldfaced words in (26a), termed “light verbs” in linguistics, serve to make verbal predicates out of nouns. It is a special feature of English that such verb-noun collocations select different light verbs, which must be memorized by learners. By comparison, Japanese uses a single light verb *suru* ‘do’ for all such expressions, as in *kyūkei-suru* ‘take a break’, *toribiki-suru* ‘make a deal’, and *undō-suru* ‘do exercises’ (similarly in Korean, where the general-purpose light verb is *bada* ‘do’). In (26b), we can see that instead of a single “and,” Mandarin speakers can choose from a number of synonymous conjunctions depending on the context. Thus, the “and” in the book title *Harry Potter and the Philosopher’s Stone* is *yǔ*, while that in the ceremony name “Parade commemorating 70th anniversary of the victories of Anti-Japanese War of the Chinese people and the World Anti-Fascist War” is *jì*.

There are also linguistic elements with a more or less even mixture of lexicality and functionality, such as numeral classifiers, which exist in a range of languages. The examples in (27) are from Mandarin Chinese and Japanese.

- (27) a. *liǎng-zhī<sub>1</sub> bǐ* ‘two-CL pen’ [Mandarin Chinese]  
           *yì-zhī<sub>2</sub> cāngshǔ* ‘one-CL hamster’<sup>12</sup>  
           *sān-zhāng zhàopiàn* ‘three-CL photograph’
- b. *nī-hon no pen* ‘two-CL GEN pen’ [Japanese]  
       *it-piki no hamusutā* ‘one-CL GEN hamster’  
       *san-mai no shashin* ‘three-CL GEN photograph’

The classifiers *zhī<sub>1</sub>*/*hon* are used for long, thin objects; *zhī<sub>2</sub>*/*biki* (the latter becomes *piki* due to a phonological process) are used for small animals; and *zhāng*/*mai* are used for thin, flat objects. In classifier languages like Chinese and Japanese, different nouns require different classifiers, but all classifiers share the same grammatical function—they all turn mass concepts into countable units. Classifiers lie somewhere near the midpoint of the continuum in Figure 1, for they are more functional than semilexical elements (with their fundamental status in the grammar being functional) and more lexical than semifunctional elements (with their idiosyncratic content being more substantive than just pragmatic conditioning).

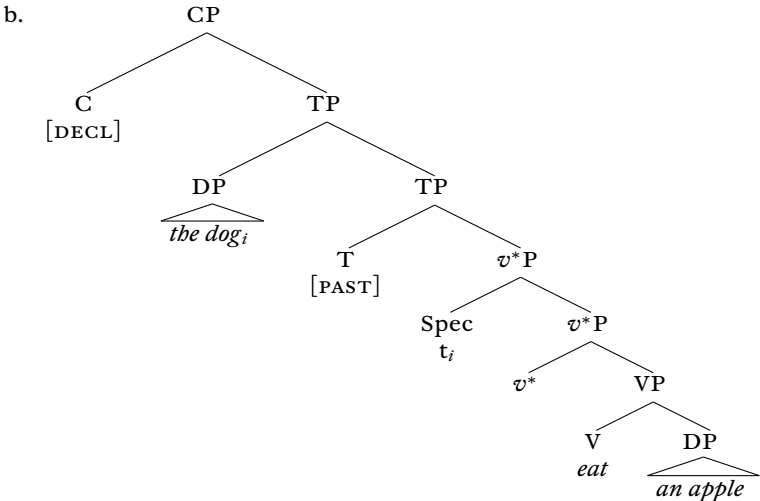
For the purpose of this article, I will simply use “semilexicality” as a cover term, without further distinguishing the fine-grained subtypes above. The phenomenon as a whole is receiving increasing attention in theoretical linguistics (see Song, 2021 for a typological discussion).

### 3.2. Generalized Root Syntax

The semilexicality phenomenon is a challenge for formal syntax, where syntactic categories are either lexical or functional, with no third possibility. This has to do with the way in which the lexicon and the syntax are theoretically connected. Simply put, syntactic derivation in the Minimalist Program starts with a lexical base, to which multiple layers of functional extension are added. Take the simple sentence in (28a) for example. Its syntactic structure is diagrammatically represented in (28b) (with some simplification for expository convenience). The tree diagram can be read from the top down as follows: “The sentence in (28a) is a CP consisting a functional head C and a TP complement selected by C; TP consists of ....”

12. The two classifiers *zhī<sub>1</sub>* and *zhī<sub>2</sub>* are etymologically unrelated and also written differently in the Chinese script, respectively as 枝 and 隻.

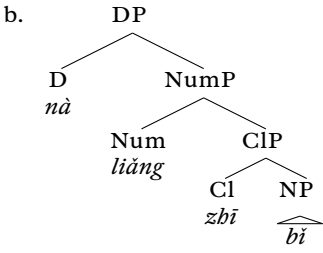
(28) a. The dog ate an apple.



The syntactic tree in (28b) shows the formal derivation of a clause, which consists of a single lexical category V plus three functional categories  $v^*$ , T, and C. These respectively serve to specify the agentive subject (i.e., the doer),<sup>13</sup> the tense (past), and the type of the clause (declarative). The lexical category V itself, on the other hand, is only responsible for introducing the core predicate (an eating activity) and its direct object (*an apple*). Leaving many technical details aside (e.g., the Spec node and the two DP triangles), we should notice that a syntactic category or “head” in the tree is either lexical or functional. There is simply no other possibility.

Now, let’s turn to a phrase with a typical semilexical element, as in (29).

(29) a. *nà liǎng zhī bǐ* [Mandarin Chinese]  
 those two CL pen  
 ‘those two pens’

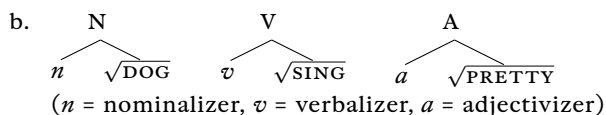


13. The subject subsequently moves to a higher position by transformation, which is conveniently indicated by a  $t$  (for “trace”) and an index  $i$  in (28b).

The Mandarin Chinese phrase in (29a) has one lexical category N, which forms its base, plus three functional categories: Cl (for the classifier), Num (for the numeral), and D (for the determiner). Crucially, the Cl head is *functional* in the syntactic system despite the semilexical nature of actual classifiers. It is impossible to reflect the semilexicality at the categorial level of the formal representation.

Linguists have noticed the above theoretical problem and also made attempts to bypass it. A representative solution, which has been independently proposed several times in recent years, is to resort to a *root-based analysis* (see, e.g., Acedo-Matellán and Real-Puigdollers, 2019, Song, 2019, and Pots, 2020). Root is a notion from an influential offshoot of generative syntax known as Distributed Morphology (henceforth DM; Halle and Marantz, 1993 et seq.; see Harley and Noyer, 1999 for a concise introduction), which treats word structure as syntactic structure and assumes a single generative engine for human language (i.e., the syntax). In DM, the root formalizes the idea of a categoryless (aka acategorial), purely lexical element, which does not even have a major part of speech. A key hypothesis of DM is that the atoms of syntactic derivation are acategorial roots (aka l-morphemes) and purely functional categories (aka f-morphemes) instead of ready-made words. On this hypothesis, what used to be considered minimal syntactic objects, most typically bare words of the major parts of speech, are given a further layer of subatomic analysis, as in (30).

(30) a. dog, sing, pretty



The three roots in (30), which are typeset in small capital letters and put under a square root symbol, are void of categorial information. They only get “categorized” by being merged with a special functional head, called a *categorizer*. Thus, the  $n\text{-}\sqrt{\text{DOG}}$  merger yields a noun *dog* based on the root  $\sqrt{\text{DOG}}$ . If we merge the same root with a different categorizer, we may get a different word of a different category. For instance, the  $v\text{-}\sqrt{\text{DOG}}$  merger yields a verb meaning “to follow very closely” or “to ask constantly.” Of course, which categorizer-root merger yields what word—or whether it corresponds to an existing word at all—is a matter of language-specific lexicalization. Thus, while the root  $\sqrt{\text{DOG}}$  is the base of both a noun and a verb (and apparently also an adjective, as in *dog French*), the root  $\sqrt{\text{BOY}}$  is only the base of a noun in current English—though a verb or an adjective *boy* is a theoretically possible word and may well be coined. The DM categorization schema just formally represents

the intuition that each content word (in a given context) has a specific syntactic category plus some idiosyncratic substantive information.

In standard DM, the root categorization tool is only reserved for content words. However, as Acedo-Matellán and Real-Puigdollers (2019), Song (2019), and Pots (2020) among others argue, it may be applied to semilexical words too. The logic is simple: when the categorizer is not a major-part-of-speech f-morpheme but an ordinary functional category, its merger with a root essentially yields a function word with some idiosyncratic content (contingent on language-specific lexicalization). Song (2019) explicitly distinguishes this extended use of the DM tool from its original use by calling the former Generalized Root Syntax.<sup>14</sup> See (31) for an illustration.

- (31) a. *yǔ* ‘and (formal/literary)’ [Mandarin Chinese]  
       *zhī* ‘classifier for long, thin objects’
- b.        Conj                      Cl  
           /        \                      /        \  
       Conj      $\sqrt{Y\check{U}}$                 Cl      $\sqrt{ZH\bar{I}}$

As we can see, Conj and Cl are both normal functional heads, but when they are respectively supported by the roots  $\sqrt{Y\check{U}}$  and  $\sqrt{ZH\bar{I}}$ , we get a conjunction and a classifier with additional idiosyncratic content (which, in the latter case, is just the usually understood idiosyncrasy of the classifier). These roots can in theory merge with other functional categories to yield other words, and this is indeed the case. Thus,  $\sqrt{Y\check{U}}$  can also be categorized into a preposition meaning “with” (32a), and  $\sqrt{ZH\bar{I}}$  can also be categorized into a verb meaning “prop up, put up” (32b).

- (32) a. *tāmen xīwàng yǔ jiārén yìqǐ guò-jíe* [Mandarin Chinese]  
       they hope with family together spend-holiday  
       ‘They hope to spend the holiday with their families.’
- b. *máfan nǐ bǎ sǎn zhī-kāi yíxià*  
       bother you DISP umbrella put.up a.bit  
       ‘Could you please put up the umbrella for me?’

### 3.3. Sentence-final emojis, formally

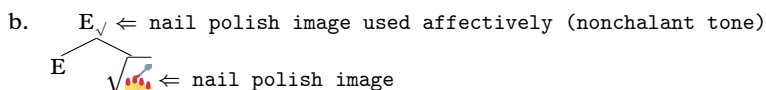
The same analytical method can be applied to SFEs. That is, we can separate their shared function (i.e., marking speaker affects) from their specific content (i.e., the affects) by encoding the former in a functional

14. Borer (2013) has a similar proposal in a different (non-DM) theoretical setting.

head, call it E (for “emotion”), and the latter, in an acategorical root. What is unique about the root of an SFE is that it is visual-digital instead of verbal-linguistic.

Since CMC is not confined by the conventional modalities of communication (e.g., oral-auditory, visual-manual), the theoretical space of roots—and thereby words in a broad sense—can be extremely large. The digital modality makes available a wide variety of elements (e.g., icons, pictures, GIFs) that may be readily recycled for communicative purposes. Since such recycled visual elements each associate a form with a (contextualized) meaning, their role in CMC is just like that of words in face-to-face (or voice-to-voice) communication—though clearly the shape of words is much more versatile in CMC. We can call the pre-recycling visual elements *digital roots* and call the communicative recycling procedure itself *digital categorization*. I illustrate this procedure in (33) with the nonchalant-tone SFE 🙄 (the example sentence is repeated from (9)). The subscript  $\sqrt{\phantom{x}}$  notation in (32b) indicates that the abstract category E is now supported by the idiosyncratic information of a root.

- (33) a. had ‘hug’ been a little more second longer, she would’ve elbowed one of these queens out. just saying 🙄 (Twitter)

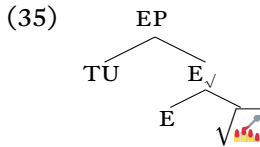


The image 🙄 itself does not necessarily denote nonchalance. At face value, it is just a nail polish icon, which may well just denote a nail-polishing activity in a different context, as in (34).

- (34) Enroll for various nail courses at Riva and pursue your dreams of becoming a nail 🙄 technician. (Twitter)

What triggers the nonchalance reading of 🙄 in examples like (33a), therefore, is the affective categorial context—or in formal linguistic terms, the functional category E. And that reading itself is a result of conventionalization, just like the meaning of any content word or idiom. In defense of Root Syntax, Marantz (1995) famously asserted that *cat* was a phrasal idiom. By the same token, we can say that each affective emoji is a tiny idiom in the CMC lexicon, because we cannot predict its affective reading with full confidence (even for simple smileys like 😊, which is more passive-aggressive than friendly in current usage) but must learn it as we learn any other new word.

Following the categorization step in (33b), the root-supported  $E_{\sqrt{\phantom{x}}}$  can project its own phrase structure like any other functional category can. This gives us the structure in (35), where TU stands for “text unit.”



The root-supported  $E_{\checkmark}$  merges with the text unit it accompanies and labels the product of this merger EP. In other words, E functions like an emotional wrapper around the text unit it accompanies.

The root-based syntactic analysis above makes several immediate predictions about the behavior of affective emojis, which exactly correspond to what we observed in Section 2. First, since the affective meaning triggered by the  $E_{\checkmark}$  merger is a result of language-specific conventionalization, the same emoji form may have different meanings in different languages/cultures or for people of different generations. In other words, emojis are not a universal language, contrary to a popular impression. The above-mentioned simple smiley 😊 is a good example of cross-generational variation. The shift in its affective meaning is similar to that in the meanings of content words like *awful* ‘impressive→extremely bad’ and *gay* ‘joyous→homosexual’. An example of cross-cultural variation is the aforementioned Weibo emoji 🙄, which is popularly used in China to express an onlooker attitude but does not have this usage in other cultures. Similarly, the dog-head emoji 🐶 (popularly named “doge”), which does not exist in Unicode but does on a number of Chinese platforms (e.g., 🐶 on WeChat, 🐶 on Douyin), has more or less become *the* emoji for sarcasm in China, as in (36).

- (36) A Weibo user posted that they had brought a lot of food to the quarantine hotel, and someone replied:



zěnmé méi      bǎ kōngqì zháguō dàishàng 🐶 [Mandarin Chinese]  
 how not.have DISP air fryer bring.along SFE  
 ‘How come you haven’t brought along your air fryer? (sarcastic tone)’  
 (Weibo)

With the dog-head emoji, it is clear to Chinese speakers that the question is not genuine but sarcastic (though not really hostile).

The second prediction of the analysis is that affective emojis are peripheral word-order-wise. They can be either to the left or the right of the text unit they accompany, but cannot be in its middle. Formally speaking, this is because the position of  $E_{\checkmark}$  is outside of the TU position in (35). The conversion of hierarchical syntactic structures to linear strings is rule-based, and there are only two linearization possibilities for the tree in (35):  $TU \prec E_{\checkmark}$  or  $E_{\checkmark} \prec TU$ . This means that there can be truly sentence-initial affective emojis beyond the marginal cases in Section 2.5, which is a point that needs further attestation. For now, we can probably explain the predominantly sentence-final positioning






of affective emojis by the content-before-emotion communicative habit of Internet users and the left-to-right directionality of the scripts in our data. This means that in languages with right-to-left scripts, affective emojis will show up to the left of the text they accompany. This is indeed the case, as evidenced by the Hebrew example in (37).

- (37)  אהבכם אוהב [Hebrew]  
 etkhem ohev  
 SFE you.PL love  
 'Love you. (affectionate tone)' (Twitter)

The blue heart emoji in (37), despite its geometric initiality, is logically sentence-final. Interestingly, the translation functionality of Twitter would automatically switch the geometric positioning of emojis too when translating from Hebrew to English.

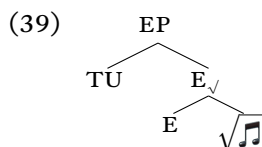
The third prediction of the root-based analysis is also about linearization. The above-mentioned two possibilities to order TU and  $E_{\sqrt{}}$  are still based on requirements of the oral-auditory modality—in particular, the requirement that linguistic structures must unfold linearly in time. However, such strict linearity is not a requirement of CMC, because the channel of externalization (i.e., the digital screen) is two-dimensional. Thus, the positioning of  $E_{\sqrt{}}$  with respect to TU ought to have more flexibility than what we have seen so far. In theory, the EP structure in (35) can be externalized in any way that does not interpolate  $E_{\sqrt{}}$  *inside* TU. Thus, we can view  $E_{\sqrt{}}$  and TU as being placed in two different layers (as in Photoshop), which may be organized in whatever way the 2D screen allows for: horizontally, vertically, or with overlay. This extended view of EP linearization makes it possible to give affective emojis and memes a unified formal analysis. See (38) for an illustration.

- (38) a.   
 is for me?
- b. 
- c. 

The three memes in (38) are respectively in English, Chinese, and Spanish, and they each externalize EP in a different way: vertically in (38a), with TU-over- $E_{\sqrt{}}$  overlay in (38b), and with  $E_{\sqrt{}}$ -over-TU overlay in (38c). Note that (38c) does not really involve interpolation despite its separation of the content of TU on two sides of the affective image, because when reading the meme, we still read the text as *Ojalá TODO vaya*

*bien* instead of *Ojalá TODO* 😬 *vaya bien*. Besides, the distributed positioning of the Spanish sentence is not just with respect to the image either, but is more exactly with respect to the entire canvas (to further use the analogy with Photoshop) and everything contained in it, as is evidenced by the larger-than-usual space between *Ojalá* and *TODO*. In other words, what we see in (38c) is a case of *geometric* rather than *logical* interpolation. The root-based analysis merely predicts the impossibility of the latter but not that of the former, for geometric positioning has more to do with graphic design than with linguistic externalization.

Last but not least, the above analytical framework allows us to further expand the scope of affective elements. The digital modality is more flexible than naturally evolved biological modalities not only in terms of image type (icons, emojis, GIFs, etc.) and linearization possibility, but also in terms of the more general “filetype” of the affective element. So far, we have limited our discussion to affectively recycled visual elements, but on the Internet, audio elements may be recycled too. This is what happens in Instagram posts or “stories” with background music. The linguistic structure of such multimedia posts is exactly the same as that of affective emojis/memes, as in (39), where I use 🎵 to denote some audio element.<sup>15</sup>



In sum, the digital modality provides a much bigger stage for the affective modification of linguistic expressions than biological modalities do, of which affective emojis are just a particular manifestation. The root-based analysis presented in this section is suitable for the affective recycling of all kinds of multimedia material.

#### 4. CMC linguistics

In Section 1, I asked two general questions: one about the cognitive nature of CMC, and the other about tools from modern linguistics that are applicable to it. My investigation of affective emojis in Sections 2–3 reveals that there is indeed some substantial cognitive difference between

15. More often than not, Instagram posts with background music also have background images. On the current analysis, this requires the root part of the structure to be a multimedia compound, which is similar to the situation with emoji sequences we have seen on p. 165.

oral languages and CMC. The difference mainly lies in the nonbiological nature of the digital modality, whose flexibility and extensibility are far beyond the capacity of naturally evolved modalities of communication. We have discussed visual and audio affective elements in this article, but as newer technologies arise, there will certainly come newer types of communicative elements too, such as elements of virtual reality or the metaverse.

The unique features of CMC requires us to rethink the relation between language and writing/typing in the 21st century. CMC is clearly still built on conventional linguistic content, either written/typed or spoken/recorded. But the ever-increasing information processing and transmission power of the computer enables users to further modify the linguistic content in unprecedented ways. It is such computer-mediated modification that requires linguists' careful investigation. The reason is that such modification counts as an "interface" issue of the digital modality.

In generative linguistics, especially in the Minimalist Program, interface legibility conditions are taken to be a major driving force and gauge of success for linguistic theory. These are conditions that a generative theory of human language must meet to ensure that the structures it generates are legible in the cognitive systems that the language faculty interfaces with. For instance, to make sure that linguistic structures are legible by the sensorimotor system, some algorithm must apply to convert them into linear strings. An influential proposal in this regard is Kayne's (1994) Linear Correspondence Axiom. On the other hand, to make sure that the linguistic structures are legible by the conceptual-intentional system, some operations must apply to remove uninterpretable features from them, such as features of grammatical case (e.g., accusative) and agreement (e.g., first-person singular). Quite a few key operations of the Minimalist Program (e.g., Agree, Delete) are motivated by this legibility condition.

Given the fundamental significance of interface conditions, linguists must ask themselves whether the same conditions still apply in the case of CMC. This is a legitimate question, because each interface presumably has its own legibility conditions. My case study in this article demonstrates that the syntax-pragmatics interface is strongly influenced by the change of modality, because CMC makes available a myriad of communicative elements (e.g., affective emojis) that take effect at the pragmatic level. Beyond the immediate scope of this article, however, I think the big-picture question we need to ask is:

- How must linguistic theory adapt itself to the cyber-digital interface?

By the cyber-digital (henceforth C-D) interface, I mean the interface between the language faculty and the computer-and-network system that CMC relies on. Note that this interface is an unusual one from a linguistic perspective, because while all other linguistic interfaces are within

the confines of the mind, the C-D interface is not—unless the computer is viewed as an extension of the human mind. The unusualness of the C-D interface means that to answer the question above, we must first answer the question below:

- How likely is it for the cyber-digital system to replace the sensori-motor system as an alternative modality of language externalization in the human world?

As things currently stand (in the early 21st century), the likelihood is quite small. But if there comes a day when the answer to the last question becomes a positive Yes (i.e., when cyborgs no longer only exist in fiction), then CMC linguistics should definitely become an official branch of linguistics, even if grapholinguistics still remains marginal.

As far as I am concerned, until the legibility conditions of the C-D interface are ascertained, perhaps the safest theoretical linguistic tools to use in the study of CMC—or more exactly the CMC-specific part of CMC data (e.g., emojis)—are none other than the most basic ones—those that are *not* designed to meet the generativists' interface conditions but are independently needed by any adequate theory of human language. In particular, I can think of the following three tools, the first two of which I have already used in my case study of emojis:

1. The basic combinatorial operation that builds complex linguistic units out of simpler ones: This operation lives under various names in different theoretical frameworks. It is called “Merge” in the Minimalist Program, which is formally just set formation:  $\text{Merge}(A, B) = \{A, B\}$ .
2. The recycling of existing materials for new purposes: This is essentially what Generalized Root Syntax is about, where miscellaneous root materials may be recycled to support and enrich abstract functional categories. Depending on the nature of the particular functional category, this may correspond to “grammaticalization” or “lexicalization” in traditional linguistic terminology.
3. The compositional interpretation of syntactic structures: This is what another major branch of theoretical linguistics, formal semantics, is about. Since the formal tools in compositional semantics (e.g., the lambda calculus, first-order logic) are not limited to the analysis of natural languages but are generally applicable to any symbolic system, they can certainly be used to represent the semantics of CMC data too.<sup>16</sup>

Thus, the safest tools to use in CMC linguistics, for the time being, are either tools that are not motivated by interface conditions or tools that

---

16. See Song (2022) for a compositional semantics for the emoji syntax proposed in this article.

are not designed for the analysis of natural language alone. On that note, the first two tools above are perhaps not entirely natural language-specific either but may be viewed as the manifestation of some domain-general strategies in the language domain: Merge qua set formation is obviously needed in many cognitive domains (e.g., mathematics), while the recycling of existing materials for new purposes is essentially just assigning old materials new categories, and categorization is one of the most fundamental cognitive processes underlying human intelligence, which clearly is domain-general too. On the other hand, many familiar generative linguistic tools (e.g., movement, phase-based spell-out) are not entirely safe due to their oral language-specific nature, or more generally due to their strong association with the legibility conditions of the sensorimotor interface. I have refrained from using such tools in my analysis of affective emojis.<sup>17</sup>

The “safe” nature of domain-general tools is reminiscent of what Chomsky (2005) has designated the “third factor” in language design—namely, principles that are not specific to the language faculty, such as principles of data analysis or processing and principles of structural architecture and efficient computation. According to Chomsky, such principles are not motivated by the need of the language faculty alone but are nevertheless an indispensable part of the growth of language in the individual. It seems therefore that the part of the generative linguistic tool kit suitable for research on CMC (again until its interface conditions become clear) is just the set of tools that can be cast as third-factor strategies.

## 5. Conclusion

In this article, I presented a formal linguistic study of affective emojis (aka sentence-final emojis) in CMC data and laid out some preliminary thoughts on CMC linguistics. The point of departure for my case study is the syntactic analysis of such emojis in Song (2019). While I have inherited and revised Song’s (2019) root-based analysis, I have objected to his unified treatment of sentence-final particles in oral languages and sentence-final emojis in CMC based on three arguments (§2). My revised analysis (§3) separates CMC data with affective emojis into a non-CMC-specific part (i.e., the linguistic text) and a CMC-specific part (i.e.,

---

17. I am not denying the utility of domain-specific tools in the study of CMC data in general but merely cautioning against their application in the study of the CMC-specific part thereof, such as emojis. One can certainly use operations like movement in the analysis of the linguistic expression basis of CMC data; i.e., the TU part of (35). A caveat here is that the separation of CMC data into a CMC-specific and a non-CMC-specific part might entail a more complicated (and potentially multiparty) interface relation between the various systems involved in CMC linguistics.

the emoji), with the latter functionally wrapping around the former and thereby setting a tone for it. A merit of this analysis is that it can be directly applied to other CMC-specific affective elements too, such as memes and background music. More generally, the analysis is suitable for any affective modification of linguistic expressions in the digital modality of communication. The special nature of the digital modality has nontrivial ramifications for CMC linguistics (§4). Until the legibility conditions of the cyber-digital system are ascertained, the safest linguistic tools to use in research on CMC-specific phenomena are the domain-general ones, or the ones that can be cast as Chomsky's (2005) "third factor" strategies. I will explore the legibility conditions of the C-D interface as well as the interface relation(s) in CMC linguistics in future research.

## References

- Acedo-Matellán, Víctor and Cristina Real-Puigdollers (2019). "Roots into functional nodes: Exploring locality and semi-lexicality." In: *The Linguistic Review* 36.3, pp. 411–436.
- Biberauer, Theresa and Ian Roberts, eds. (2013). *Challenges to Linearization*. Berlin: De Gruyter Mouton.
- Borer, Hagit (2013). *Taking Form*. Vol. 3. Structuring Sense. Oxford: Oxford University Press.
- Carey, John (1980). "Paralanguage in computer mediated communication." In: *Proceedings of the 18th Annual Meeting of the Association for Computational Linguistics*, pp. 67–69.
- Chao, Yuan-Ren (1968). *A Grammar of Spoken Chinese*. Berkeley: University of California Press.
- Cheng, Siu-Pong and Sze-Wing Tang (2022). "Syntax of Sentence-final Particles in Chinese." In: *The Cambridge Handbook of Chinese Linguistics*. Ed. by Chu-Ren Huang et al. Cambridge Handbooks in Language and Linguistics. Cambridge: Cambridge University Press, pp. 578–596.
- Chomsky, Noam (1957). *Syntactic structure*. The Hague: Mouton.
- (1959). "On certain formal properties of grammars." In: *Information and Control* 2.2, pp. 137–167.
- (1995). *The Minimalist Program*. Cambridge MA: MIT Press.
- (2001). "Derivation by phase." In: *Ken Hale: A Life in Language*. Ed. by Michael Kenstowicz. Oxford: Oxford University Press, pp. 1–52.
- (2005). "Three factors in language design." In: *Linguistic Inquiry* 36.1, pp. 1–22.
- Durrell, Martin (2021). *Hammer's German Grammar and Usage*. 7th ed. Routledge Reference Grammars. Routledge, Taylor et Francis Group.

- Fehringer, Carol and Leonie Cornips (2019). "The Use of Modal Particles in Netherlandic and Belgian Dutch Imperatives." In: *Journal of Germanic Linguistics* 31.4, pp. 323–362.
- Gawne, Lauren and Gretchen McCulloch (2019). "Emoji as digital gestures." In: *Language Internet* 17.2. [https://www.languageatinternet.org/articles/2019/gawne/index\\_html](https://www.languageatinternet.org/articles/2019/gawne/index_html).
- Grosz, Patrick Georg et al. (2021). "A semantics of face emoji in discourse." Manuscript, to appear in *Linguistics & Philosophy*. <https://ling.auf.net/lingbuzz/005981>.
- Halle, Morris and Alec Marantz (1993). "Distributed Morphology and the pieces of inflection." In: *Essays in Linguistics in Honor of Sylvain Bromberger*. Ed. by Ken Hale and S. Jay Keyser. The View from Building 20. Cambridge MA: MIT Press, pp. 111–176.
- Haralambous, Yannis (2020). "Grapholinguistics, T<sub>E</sub>X, and a June 2020 conference." In: *TUGboat* 41.1, pp. 12–19.
- Harley, Heidi and Rolf Noyer (1999). "Distributed Morphology." In: *Glott International* 4.4, pp. 3–9.
- Kayne, Richard (1994). *The Antisymmetry of Syntax*. Cambridge MA: MIT Press.
- Li, Charles and Sandra Thompson (1981). *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.
- Maier, Emar (2021). "Emojis as pictures." Manuscript, to appear in *Ergo*. <https://ling.auf.net/lingbuzz/006025>.
- Marantz, Alec (1995). "Cat as a phrasal idiom: Consequences of late insertion in Distributed Morphology." Manuscript, Massachusetts Institute of Technology.
- Morita, Emi (2018). "Sentence-final Particles." In: *The Cambridge Handbook of Japanese Linguistics*. Ed. by YokoEditor Hasegawa. Cambridge Handbooks in Language and Linguistics. Cambridge: Cambridge University Press, pp. 587–607.
- Paul, Waltraud (2014). "Why particles are not particular: Sentence-final particles in Chinese as heads of a split CP." In: *Studia Linguistica* 68.1, pp. 77–115.
- Pierini, Francesco (2021). "Emojis and gestures: A new typology." In: *Proceedings of Sinn und Bedeutung*. Vol. 25, pp. 720–732.
- Pots, Cora (2020). "Roots in Progress: Semi-lexicity in the Dutch and Afrikaans verbal domain." PhD thesis. KU Leuven.
- Potts, Christopher (2005). *The logic of conventional implicatures*. Oxford: Oxford University Press.
- Saussure, Ferdinand de (2011). *Course in General Linguistics*. illustrated, revised. Ed. by P. Meisel and H. Saussy. Translated by Wade Baskin. Columbia University Press.
- Song, Chenchen (2019). "On the formal flexibility of syntactic categories." PhD thesis. University of Cambridge.

- Song, Chenchen (2021). "A typology of semilexicality and the locus of grammatical variation." Talk at the 9th International Conference on Formal Linguistics (ICFL9), Nov 5–7, Fudan University (online) <https://www.juliosong.com/doc/Song2021ICFL9.pdf>.
- (2022). "Sentence-final particle vs. sentence-final emoji: The syntax-pragmatics interface in the era of CMC (extended version)." SyntaxLab talk, Jun 28, University of Cambridge (online) <https://www.juliosong.com/doc/Song2022SynLabJun.pdf>.
- Sun, Xixin (1999). *Interjections in Modern Chinese*. Beijing: Language and Culture Press.



# The Rosetta Stone Squandered: Decipherment's Twelve-Year Gap and the Fate of J.D. Åkerblad

Daniel Harbour

*Abstract.* Just three years after its much feted discovery, the Rosetta Stone fell into a period of sustained neglect. Two partial decipherments had been made of its demotic text but its hieroglyphic inscription had barely been investigated. I examine the reasons behind this fall from grace and argue that J.D. Åkerblad, the author of the more penetrating demotic study, could have made significant inroads into the decipherment of Egyptian hieroglyphs. His results would have presaged by twenty years results of the eventual decipherer of the script, Jean-François Champollion. This would have changed, with untellable consequences, the intellectual space in which Champollion and his main rival, Thomas Young, worked. The study's conclusions highlight the centrality of decipherment to philology/grapholinguistics and the importance, both to research and to researchers, of properly functioning academic institutions.

## 1. Introduction

The Rosetta Stone, 'that long-desired monument ... which will probably lead us one day to a knowledge of ancient Egyptian writing', fell into a period of scholarly neglect a mere three years after its discovery. Given the fanfare, and warfare, that surrounded the find—Napoleon's prize, then King George's—this sudden desuetude is astonishing. It was, furthermore, unwarranted. Johan David Åkerblad, quoted above (Åkerblad 1802b, p. 494), had made significant inroads into the Stone's demotic after just two months' study and was beginning to cast his eye on its hieroglyphs. Then adverse professional, political, and personal circumstances began to overtake him. This article revisits research on the Stone as it stood at the end of 1802 and demonstrates that the methods

---

I am grateful to Robert Crellin and Amalia Gnanadesikan for help and feedback, and to the organisers of /gʁafematik/ 2022 for the opportunity to present this work.

---

Daniel Harbour  0000-0002-6335-0097

Department of Linguistics (SLLF), Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom.

E-mail: d.harbour@qmul.ac.uk

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 193–217. <https://doi.org/10.36824/2022-graf-harb>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

Åkerblad deployed on its demotic were apt to yield further significant insights into its hieroglyphs. These could have brought key elements of their decipherment forward by as much as twenty years.

Decipherment is not a done-and-dusted discipline. Work on oracle bone Chinese, Aztec, and Linear Elamite has made major progress just in the past few years (Jiǎng 2018, Whittaker 2021, Desset et al. 2022). At the same time, pseudodecipherments continue to be proposed, including for the Rosetta Stone.<sup>1</sup> In this context, progress in decipherment consists not only in unlocking still mysterious scripts and signs but in revisiting past work. Just as mathematicians seek alternative proofs of established results, so can decipherers demonstrate the soundness of their tools by showing that paths not taken converge on the same end. That the same simple methods yield success in decipherment after decipherment points to an underlying unity in the nature of writing systems. This finding itself deserves a significant place in the emerging field of philography, which studies linguistic cognition as it is embodied in writing systems.

Åkerblad's fate and the twelve-year gap in Rosetta Stone research are two sides of the same case study in the importance of properly appreciating the power of the decipherer's tools. In 1799, the Rosetta Stone became the first ever monument to offer scholars a roughly parallel text in a known language, Greek, and an Ancient Egyptian script—two, in fact.<sup>2</sup> It consisted of fourteen lines of hieroglyphs, mostly incomplete, along with thirty-two lines of demotic and fifty-four of Greek, mostly complete. This immediately raised hopes of traction on decipherment. In 1802, and despite France's loss of the Stone to Britain, two scholars based in Paris, the French orientalist Silvestre de Sacy and Åkerblad, his Swedish diplomat student, published studies of the demotic inscription. Åkerblad's was clearly the better, correcting several errors in Silvestre de Sacy's and pushing substantially further into the script. Yet the field then stalled, not regaining momentum until the mid 1810s when Thomas Young and Jean-François Champollion, armed with a substantially larger supply of materials—in particular, the *Description de l'Égypte* and later the Philae obelisk (Commission des sciences et arts d'Égypte

---

1. In 2010, the Macedonian Academy of Sciences and Arts published a purported proof that the Rosetta Stone demotic records a Slavic language, closely affiliated, needless to say, to Macedonian itself. The work made no effort to engage with two centuries' worth of discoveries about demotic and its relationship to hieroglyphs and hieratic, for which reason I decline to cite it directly. It can be found at <http://manu.edu.mk/contributions/NMBSci/vol31p1.html>.

2. For studies of the Stone, its history, and its role in decipherment, see Parkinson 1999, Solé and Valbelle 1999, Adkins and Adkins 2000, Ray 2007, and Robinson 2012, amongst others.

1809–1822, Bankes 1821)—made the crucial breakthroughs on which the full decipherments of demotic and hieroglyphs ultimately rest.

Dominant narratives of the decipherment of Egyptian writing are often hostage to the charisma of clear, iconic hieroglyphs over squiggly, indistinct demotic and to the Romantic ideal of the lone genius over a field of complex, collaborative rivalry. Both factors lead to a focus on Champollion (who, in all fairness, does deserve substantial attention, just not the monopoly that some accounts afford him). Corrections to this partial view of history are usually framed in terms of an exaggerated Anglo-French rivalry between Young and Champollion, which was carried on by their fellow countrymen, and others, long after the two decipherers had largely healed their rift and, indeed, died.

The loser in all of this is Åkerblad. Though he received many accolades following his 1802c publication, his neglect also began at that time and was rooted in the gatekeepers of the institutions of orientalism that were then being erected. Yet the methods that he deployed on the demotic of the Rosetta Stone were readily applicable to its hieroglyphs. They could never have led to a full decipherment. However, they could have advanced key findings and reshaped hypotheses that were ambient when Young and Champollion joined the fray. Subsequent history might have been very different.

To make this case, section 2 presents Åkerblad's methods and results, emphasising in particular his willingness to go beyond the demotic text, repairing the Greek and venturing into the hieroglyphs. Section 3 briefly outlines why Åkerblad did not go on to further study of the Stone. Section 4 then examines what he might have discovered, had he had the opportunity to carry his investigation further into the hieroglyphic text. In particular, I argue that he could have partially deciphered the contents of the Stone's two hieroglyphic cartouches, identifying the names and epithets they contain and gaining insight into the extent of phonographic writing for foreign names, Egyptian names, and normal Egyptian words. Finally, section 5 shows that these results would have impacted on the field in at least five ways, confirming hypotheses of previous work and preempting later results of Silvestre de Sacy, Young, and Champollion.

## 2. Åkerblad: Method and Results

My claims about Åkerblad's potential contributions to the study of hieroglyphs are based on the expertise in Coptic that won him access to the Rosetta Stone lithographs (section 2.1), the methods that he deployed in his study of demotic and the results he drew from them (section 2.2), and his imaginative ability to leverage small finds to push beyond the confines of the demotic text (section 2.3).

## 2.1. Access

Johan David Åkerblad came to study the Rosetta Stone by a rather circuitous route. An orientalist by inclination and training, his diplomatic posting brought him to a Paris that had shortly before received a treasure trove of Coptic and other manuscripts, Napoleon's Vatican booty. He buried himself in these, producing studies of both Phoenician and Coptic (Åkerblad 1802a,b). The latter was crucial to his gaining access to the tightly guarded Rosetta Stone lithographs.

In his wanderings through the new acquisitions at the *Bibliothèque nationale*, Åkerblad had come across a short passage of an unknown involuted script at the end of a Coptic manuscript. His powerful command of that language enabled him to see that the script was an ornately cursive form of Coptic. He quickly deciphered and translated the passage, writing his findings up as a *lettre* to Silvestre de Sacy. The latter forwarded the work, with his own brief cover letter, full of praise for Åkerblad's Coptic prowess, to the editor of the *Magasin encyclopédique*, where it was duly published. Åkerblad's letter finished with a plea to his teacher (quoted more briefly at the opening of this article; Åkerblad 1802b, p. 494):

*Je desire bien vivement, Monsieur, que l'inscription de Rosette, plus digne d'exercer la sagacité de ceux qui savent le copte, soit bientôt publiée avec vos savantes remarques. ... certes il est temps que l'on fasse connoître aux savans ce monument depuis longtemps désiré, et qui, probablement un jour, nous conduira à la connoissance de l'ancienne écriture égyptienne.*

I desire most lively, Sir, that the Rosetta inscription, much more worthy to exercise the wisdom of those who know Coptic, soon be published with your wise remarks. ... surely, it is time that long-desired monument, which will probably lead us one day to a knowledge of ancient Egyptian writing, be made known to scholars.

The study to which Åkerblad alludes had been underway for some two years at the time and the authorities above Silvestre de Sacy had begun to lose patience. The twice victorious decipherer had written his results up (Silvestre de Sacy, 1802a) with evident embarrassment at their shortcomings. Possibly to save face, he suggested that someone with greater knowledge of Coptic might be capable of greater progress. The confluence in print of these three factors—Silvestre de Sacy's lack of Coptic competence, his ample praise of Åkerblad's, and the latter's plea for scholarly access to the Rosetta inscriptions—made Åkerblad's access to the precious lithographs all but inevitable. Silvestre de Sacy could at least be comforted that he was ceding access to his own protégé; but his feelings on this point would soon change.

## 2.2. Decipherment

Åkerblad's starting point might be termed 'Leibniz's lemma'. With typical acuity, Leibniz had observed, in a letter of published in a multilingual compendium of the Lord's prayer, that proper names in bilingual inscriptions provide an entry point for decipherment (Leibniz, 1715, p. 23):

*Extant apud Palmyrenos & alibi in Syriâ, & vicinis locis complures inscriptiones antiquæ duplices, partim linguâ & Characteribus gentis, partim Græcè expressæ, quæ magnô studio ex ipsis saxis describi deberent. Inde enim fortasse constitui Alphabetum posset, & linguæ indoles tandem cognosci, cum Græca versio adsit, & Nomina Propria interveniant quorum eadem ferè in Patrio & Græco sermone pronuntiatio erat.*

Among the Palmyrenes and elsewhere in Syria, and in the neighbouring places, there are several ancient double inscriptions, partly expressed in the language and Characters of the people, partly in Greek, which should be described with great study from the rocks themselves. For, when there is a Greek version, and Proper Names appear, the pronunciation of which was nearly the same in the Native and Greek languages, from this perhaps an Alphabet might be established, and the character of the language finally known.

Leibniz was the first to formulate this use of proper names explicitly. Earlier applications of the principle are to be found in Agustín 1587, a limited venture into Iberian, and Halley 1695, an unsuccessful attempt on Palmyrene.

Åkerblad's procedure was, first, to find repeated proper names, using the ratio of Greek versus demotic text lengths as a guide to their position. (For instance, names that appeared in lines 13 and 27 of the 54-line Greek text should be located around lines 8 and 16 of the 32-line demotic, that is, a quarter- and half-way through each.) Names found in this way served as landmarks from which to locate further names, especially nonrepeated ones. When sufficient names had been found, he set about distilling an alphabet, looking for letters shared between names that shared sounds. Finally, pushing beyond proper names, he looked for further legible words in semantically plausible contexts. Examples of his haul of names and other words are given in Fig. 1 and his resulting alphabet, in Fig. 2.<sup>3</sup>

Åkerblad's contribution was not in the originality of his method. Previous decipherers had leveraged proper names in a similar fashion: Barthélemy for Palmyrene and Phoenician (1759, 1764) and Silvestre de

---

3. Image files are drawn from public-domain copies available at <https://archive.org>, <https://biodiversitylibrary.org>, and <https://books.google.com>. I have not located copies of the *Institut d'Égypte* lithographs used by Åkerblad. Slightly anachronistically, I have used the 1803 engravings published by Society of Antiquaries (Vertue, Basire, Basire, and Basire, 1815).

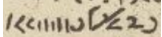
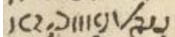

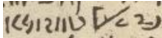
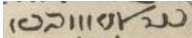
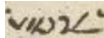
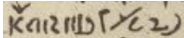

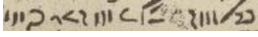
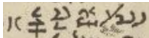
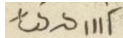
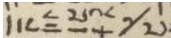
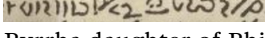

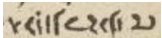
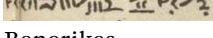
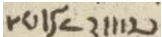
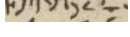
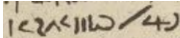
|   |   |  |
|---|---|--|
| Ptolemy   | Arsinoë   | Egypt  |
|  |  |   |
|  |  |   |
|  | Areia daughter of Diogenes  |   |
| Alexander, Alexandria   |  | Greek ('Ionian')   |
|  | Irene daughter of Ptolemy   |  |
|  |  | Osiris   |
| Aëtos   | Pyrrha daughter of Philinos   |   |
|  |  | taxes (< <i>syntaxeîs</i> )  |
|  | Benerikes   |  |
|   |  |  |


FIGURE 1. Demotic names and words isolated by Åkerblad

|                                     |                                   |
|-------------------------------------|-----------------------------------|
| α ... 2 ... 2 (α, β)                | ε ... 91 ... 91 ... 2911 ... 1911 |
| β ... 4                             | τ ... 7 ... 7 ... 7               |
| γ ... 7 (γ)                         | ρ ... 1 ... 1                     |
| δ ... 7 (δ)                         | φ ... 3 ... 3 ... 3 ... 3 mod.    |
| ε ... 1 ... 2 ... 7 (ε, ζ)          | χ ... 2                           |
| ζ ... 7                             | ψ ... 2 ... 7 (ψ)                 |
| η ... 111 ... 111 ... 111 (η, θ, ι) | ω ... (ω, ρ) ... 7 ... 7          |
| θ ... 7                             | υ ... 7 ... 7 ... 7 ... 7         |
| ι ... 5 ... 11 ... 11               | φ ... 4 ... 4                     |
| κ ... 7 ... 7 ... 7                 | χ ... 2 ... 2 ... 2               |
| λ ... 7 ... 7                       | θ ... 7                           |
| μ ... 7 ... 7 ... 7                 | ζ ... 7 ... 7 ... 7               |
| ν ... 7 ... 7 ... 7 mod.            | σ ... 7 ... 7 ... 7               |
| ξ ... 7 ... 91 ... 7 (ξ, η)         | τ ... 7                           |
| ο ... 7 ... 7 ... 7                 | ι ... 7                           |
| π ... 7 ... 7 ... 7                 | β ... 7                           |
| ρ ... 7 ... 7 ... 7                 | γ ... 7                           |

FIGURE 2. Åkerblad's demotic alphabet



Sacy for Middle Persian and Parthian (1793), studies of which Åkerblad was well aware (Åkerblad 1802c, p. 4). Indeed, Silvestre de Sacy (1802a) had attempted the same strategy during his two-year monopoly. The results were, by his own admission, meagre. Åkerblad's contributions lay in his ability to see this method through in what, it would much later emerge, was not a simple alphabetic script of the kind where Barthélemy and Silvestre de Sacy's had enjoyed success. His results extended well

beyond his alphabet and concerned both the decipherment of demotic and insight into the rest of Stone.

Some of his discoveries corrected Silvestre de Sacy's errors. The latter had arrived at the—frankly, in the context of Near Asiatic scripts, odd—notion that 'Alexander' was spelled with four initial capitals. These are in fact stacked letters, the components of which are clearly discernible, as comparison of Fig. 1 and 2 shows. Åkerblad further identified that  was 'many', not, as Silvestre de Sacy had it, the goddess 'Isis', amongst several similar results.

More impressively, Åkerblad slipped the tether of proper names precisely where it would have hobbled him. Not all proper names are alike in their propensity to match across languages. Ethnonyms in particular are prone to mismatch (as the French, German, Italian, and Polish—*allemand*, *Deutsch*, *tedesco*, *niemiecki*—and indeed English for 'German' illustrate). Åkerblad deciphered that Egyptians called their country (something akin to) *k<sup>b</sup>emi*, like the Biblical 'Land of Ham' (𐤇𐤍 *hām*), and regarded Greeks as *wynn* 'Ionians'. Both appellations depart from the Greek.

Åkerblad's discoveries confirmed that Coptic was the linguistic key to Ancient Egyptian. Georg Zoëga (1797, pp. 455, 552–553) had drawn attention to the similarity of Coptic words and the readings that Greek sources had attributed to some hieroglyphs. The sources were thin and the readings, scant, however. They amounted at best to a weak hypothesis about the relationship between Coptic and the language of Egyptian monuments. Åkerblad pinned this down by identifying *ašai* 'many' (in fact, *ʕšʕy*), which corresponds to Coptic ⲁϣⲁⲓ *ašai* 'many', and *nierfēwi* 'temples' (in fact, *irpy*), responding to Coptic ⲉⲣⲫⲉⲓ *erfei* 'temple'. He concluded (Åkerblad, 1802c, p. 40) *que la langue Copte contient les débris de l'ancien égyptien, et qu'elle doit par conséquent servir à interpréter notre inscription* 'that the Coptic language contains the remnants of ancient Egyptian, and that it must, in consequences, aid in interpreting our inscription'—an insight that would become orthodoxy (Bunsen, 1845).

Åkerblad drove the Coptic-Egyptian connection home through discoveries at the level of individual letters and their form. The Coptic alphabet is very obviously adopted from the Greek. However, it contains a handful of characters of non-Greek origin. Åkerblad showed that these stemmed from demotic. For instance, the *ʕ* of Coptic ⲁϣⲁⲓ *ašai* 'many' in the previous paragraph clearly resembles the corresponding symbol in the middle of the demotic word  *ašai* (*ʕšʕy*). Likewise, ⲥ *ē* of ΝΙϢϢΗ *nisočē* 'left' answers to the second last (second from left) symbol of  (Åkerblad, 1802c, 46, pl. 1).<sup>4</sup> The continuity both of sounds and sym-

4. Curiously, these characters retained their sinistroverse orientation when exported to a dextroverse script.

bols strongly suggests continuity of language too (though the degree of similarity between Coptic and Ancient Egyptian would remain subject to debate for some time; Bunsen 1845, Hincks 1848).

### 2.3. Beyond Decipherment

Åkerblad's results extended beyond the demotic script and into the Rosetta Stone itself. He realised that the demotic was a translation of the Greek, not the reverse. One illustrative argument concerned the transliteration of Greek names. In the demotic, male names end in *-s*, as per *pṯlwmys* 'Ptolemy', *ṯlksṯntrs* 'Alexander', and *ṯyṯtws* 'Aētos' (until further notice, I use current readings of the script, rather than Åkerblad's). No female names do (as in *ṯrsyn?* 'Arsinoë', *ṯry?* 'Areia', and *brṯn?* 'Irene'), with one exception: *brnyk?* 'Berenice'. By parity with other female names, it should have been *brnyk?*. This name occurs as the first in a long string of genitives, for which the Greek feminine singular coincides with masculine nominative *s*. The erroneous *s* is explained, Åkerblad reasoned, if the scribe was translating the text from Greek as he carved. Presented with yet another name ending in *s*, he copied that sound into the demotic. Only afterwards, for the other members of the list, did he realise that the *s* was part of the Greek declension, not the name itself. (One wonders how he or his supervisors reacted.)


Not only did Åkerblad correct the ancient scribe of the Rosetta Stone but he did the same to contemporary copyists. The inscriptions had been reproduced in Egypt by the Napoleon's *savant* army in an early attempt at (literal) lithography. The innovative technology was not entirely reliable and one error it produced occurred in the phrase 'in the priesthood of Aētos son of Aētos' (line 4 of the Greek). The copy transformed Α to Δ, warping ΑΕΤΟΥ 'of Aētos' into ΔΕ ΤΟΥ (there are no spaces in the Greek).<sup>5</sup> The phrase thus produced came in for particular comment in Ameilhon's (1803) study of the Greek and he suggested that it (or just its article) was emphatic, translating the whole as *sub pontifice Aete ... quidem*, that is, 'under the priesthood of Aētos indeed'. The demotic made clear to Åkerblad that this unusual phrase was wrong. The Egyptian text clearly repeated the name 'Aētos' (Fig. 1, bottom left). A lithographic slip had wiped a generation of Aētoses from history. Åkerblad restored them.

He used the demotic further to restore the Greek text from the missing bottom right of the Stone: 'in each of the temples of first, second, and third rank, in which a statue of the King shall be erected'. Crucially

---

5. Throughout the article, I use small caps for Rosetta Stone Greek, in imitation of the Stone itself.



for the argument that follows, Åkerblad supported this infill via a brief foray into the hieroglyphic text that surmounted the demotic. In the tally marks of , he saw a clear parallel for the rather more opaque numerals on the demotic and a completion of the sequence 'first ... second ...' partially present in the damaged Greek.

### 3. What Happened Next

Åkerblad's contribution was substantive and ambitious but its immediate aftermath was mixed. There were accolades, including honorary memberships of academic institutions, dedicated editions of journals, and triumphant epithets from their editors (Thomasson, 2014). However, there was also criticism. Friedrich Münter, the almost-decipherer of Old Persian cuneiform, rightly observed that the task was far from done (Thomasson, 2013, p. 239). (No one appreciated at the time that there was far more to demotic than its consonantal alphabet. This insight had to await Young 1830 and Brugsch 1848, amongst other works.) More influential was Silvestre de Sacy's reaction. Despite his fustian endorsement of several of Åkerblad's key points, he rather pointedly withheld endorsement from others. Even without mounting a cogent case against them, his stature was such that the effect was chilling. Writing to Young thirteen years later, Åkerblad explained (Leitch 1855, p. 31; '[perhaps]' is his addition) that:

as I had not the good fortune to satisfy the mind of the learned orientalist, to whom the letter was addressed, who formally declared, that '[perhaps] some remaining attachment to the ideas which he had himself advanced, embarrassed his opinion, and prevented his full conviction' of the truth of my interpretation, I felt no further inclination to continue an investigation, in which nobody would have been interested, after such a declaration from one of the most learned men in France. I was besides at that time intrusted [sic.] with a diplomatic commission, at first in Holland, and then in France, which made me abandon almost entirely all further inquiry respecting the Inscription of Rosetta.

This quotation makes clear that Silvestre de Sacy's reaction was not the only factor that diverted Åkerblad from further investigation. His career as a diplomat in the service of Sweden, a position that ill-fitted his antiroyalist convictions and that he would eventually abandon, preferring Rome to orders to return home, also intervened. Deprived of the Rosetta Stone lithographs, he nonetheless continued to contribute to orientalist linguistics, specifically Arabic and Samaritan, via the Leiden library (Thomasson, 2013, p. 242).

Silvestre de Sacy's role in depriving the field of Egyptology of what might have become one of its leading lights appears more deliberate

than accidental. Åkerblad was one of three contemporary scholars who studied Egyptian place names in the hope of further insight into the ancient language of the country. The others were Etienne Quatremère and the eventual decipherer of hieroglyphs, Jean-François Champollion. Of the three, Silvestre de Sacy favoured his fellow royalist Quatremère. Champollion rushed his study to print once news of Quatremère's broke. Åkerblad's study, by contrast, had the misfortune to fall into Silvestre de Sacy's hands, where it languished in a deliberate act of what Young would have called 'literary injustice':

*M. Étienne Quatremère ... traitait avec beaucoup d'érudition le même sujet ... Je crus alors inutile de donner aucune publicité au mémoire de M. Åkerblad, qui me reprocha même, non sans quelque fondement, d'avoir été cause qu'il avait été prévenu par M. Quatremère. ... j'avais fini par perdre de vue son mémoire manuscrit, qui était resté entre mes mains. Une circonstance dont il est inutile de parler m'en ayant rappelé le souvenir, j'ai cru convenable d'en faire jouir le public, et de réparer ainsi le tort involontaire dont je m'étais rendu coupable envers l'auteur.*

M. Étienne Quatremère ... treated the same subject with much erudition ... I thought it then useless to give any publicity to M. Åkerblad's memoir, who yet reproached me, not without some foundation, for having been the cause of his preemption by M. Quatremère. ... I ended up losing sight of his manuscript, which remained in my hands. Having been reminded of it by a circumstance of which it is useless to speak, I thought fit to offer it now to the public, and thus to repair the involuntary wrong of which I had been guilty towards its author.

Silvestre de Sacy wrote these words in 1834. Åkerblad had died in 1819. His manuscript, the first of two intended parts, had been completed almost a decade before (Åkerblad, 1834 [1810], p. 435). The question that the passage and its double use of 'useless' rather raises, is (as Thomasson 2014, p. 285 observes), "Useless for whom?"

Åkerblad's continued study of Coptic after 1802 and his work on Arabic and Samaritan make a compelling case that he would have continued with his investigations of the Rosetta Stone, had his professional and personal circumstances been different.

#### 4. What Might Have Happened

In the closing paragraph of his *Lettre*, Åkerblad remarks that *jusqu'à présent je n'ai eu le temps d'examiner que fort légèrement [la partie hiéroglyphique]* 'I have had time to examine but most lightly [the hieroglyphic part]' of the inscription (1802, p. 63). What would have happened had he had time, inclination, and encouragement to apply his methods and mind to the Rosetta Stone hieroglyphs?

#### 4.1. Finding Ptolemy

The obvious starting point for any attempt on the Rosetta Stone hieroglyphs are its cartouches. In a sea of unfamiliar symbols, these ringed and repeated portions of text stand out. The surviving fragments of the hieroglyphic inscription contain two distinct cartouches, one short, one long. As the right alignment in Fig. 3 highlights, the short one is the initial portion of the long. The long cartouche occurs three times: near the end of the inscription (end of line 14), at the start of line 12, and around the middle of line 6. The short one occurs twice, both times on line 6, shortly after the first long cartouche. Just as these double- and triple-repeated sequences seem to be highlighted to draw the reader's eye, so do they draw the decipherer's.

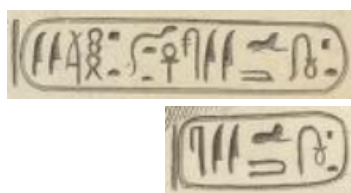


FIGURE 3. The long cartouche of the Rosetta Stone subsumes the short one

Zoëga's study of hieroglyphs, a major component of his magnum opus on obelisks, which appeared just before the discovery of the Rosetta Stone, established two key results relevant to these cartouches. First, he had used repeated phrases, occurring sometimes with and sometimes without line breaks, to establish the direction in which hieroglyphs were read, namely, against the direction in which figures in profile face (rightwards reading of leftward-facing fragments, leftwards reading of rightward-facing ones). Åkerblad would have been able to confirm this direction of reading—or arrive at it independently—based on the sinistreverse order of 'first, second, third' 𓆎𓆎𓆎.

Second, Zoëga and his predecessor Anne-Claude-Philippe de Tubières-Grimoard de Pestels de Levis Caylus had hypothesised convergently as to the function of cartouches, three decades apart. Caylus, or rather his assistant Barthélemy, reasoned by way of captioned portraits and the formulaic structure of obelisks (1762, p. 79):

*Je pense que ... ces hiéroglyphes ... sont réunis dans des ovaux ou des quarrés, pour représenter peut-être des noms de Rois & de Dieux. C'est ainsi que sur la bande inférieure de la Table Isiaque trois Figures principales sont accompagnées d'inscription hiéroglyphiques, renfermées dans de petites tables de différentes formes; c'est ainsi que*

*sur chaque obélisque les hiéroglyphes renfermés dans des ovales, sont communément distingués des hiéroglyphes que contiennent les ovales des autres obélisques; ....*

I think that ... these hieroglyphs ... are gathered in ovals or squares, to represent perhaps the names of Kings and Gods. It is thus that on the lower band of the [Bembine] Table of Isis, three main Figures are accompanied by hieroglyphic inscriptions, enclosed in small tables of different shapes; it is thus that on each obelisk the hieroglyphs enclosed in ovals are commonly distinguished from the hieroglyphs contained in the ovals of the other obelisks; ....

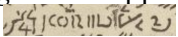
Zoëga by contrast compared the in-cartouche text with its surrounds (1797, pp. 465–466):

*Conspiciuntur autem passim in Aegyptiis monumentis schemata quaedam ovata sive elliptica planae basi insidentia, quae emphatica ratione includunt certa notarum syntagmata, sive ad propria personarum nomina exprimenda sive ad sacratiores formulas designandas. ...; uti nec illa syntagmata in alio loco inveniuntur, quae sunt ovatis schematibus inclusa.*

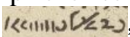
In Egyptian monuments, certain schemata are everywhere seen set in an oval or flat elliptical base, which, by way of emphasis, include expressions corresponding to proper names of persons or more sacred formulae. ...; nor are those expressions which are included in oval shapes found elsewhere [on the monuments].

Caylus' reasoning was inapplicable to the Rosetta Stone but Åkerblad could readily have checked that contents of his cartouches were indeed confined to quarters as Zoëga had specified. The question then was which names and/or sacred formulae the Rosetta Stone cartouches were likely to contain.

Common sense says to look for 'Ptolemy' on that which is Ptolemy's. This is precisely how Åkerblad had begun his investigation of the demotic. If the correct approach, then the short cartouche would be 'Ptolemy' with zero or more epithets and the long cartouche, 'Ptolemy' with at least one epithet more.

The reasoning by ratios that Åkerblad had used to locate names in the demotic confirms these associations. The first long cartouche, around the middle of line 6, is three times as far from the final long cartouche as the second one is. The final cartouche occurs near the end of the hieroglyphic text, hence almost at the end of line 14, whereas the second occurs near the start of line 12. Thus the first and second cartouches are at distances of eight and half and just under three lines from the last one. This gives an approximate ratio of three to one. In the demotic, the phrase  (epithetised 'Ptolemy') occurs with the same ratio of distances, ten lines and three-and-a-bit lines, from the end (more specifically, towards the end of line 22 and two thirds of the way along line 29 in a 32-line text).

Distribution of the remaining cartouches further supports the identification of 'Ptolemy'. Both short cartouches occur in quick succession

after the first long one. This points the decipherer to the start of line 23, where there are indeed two occurrences of , the very sequence that Åkerblad had deciphered as 'Ptolemy'. Figure 4 aligns this and the previous demotic phrase in the same fashion as Fig. 3, showing that the demotic, like the hieroglyphs, share their initial (right) segments. Thus, the long-short-short pattern of the cartouches on line 6 of the hieroglyphic text matches that of 'Ptolemy' with and without epithets on lines 22–23 of the demotic.

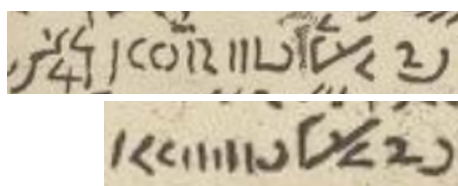
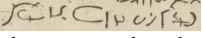
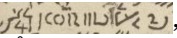



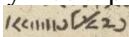


FIGURE 4. A second superstring/substring pair: demotic correspondents of the long and short cartouches

There is a problem, however. The demotic ends with  instead of the epithetised 'Ptolemy' , which a completely parallel text would lead us to expect. (Åkerblad would have been primed for such mismatches, having discovered several between the demotic and Greek.) Only the very first character and the last few of the two images coincide. Åkerblad might have seen the initial character as the opening curve of the cartouche. Instead, he misread this 'punctuation mark' as Coptic *m*, an article. Notwithstanding, he would not have been at an impasse here. The opening portion  of what should be the final cartouche is ubiquitous in the demotic text, with well over two dozen occurrences. Its recurrence in quick succession on line 1 of the demotic suggests the reading 'king', given the repetitions of the root *ΒΑΣΙΛ-* on line 1 of the Greek: *ΒΑΣΙΛΕΥΟΝΤΟΣ ΤΟΥ ΝΕΟΥ ΚΑΙ ΠΑΡΑΛΑΒΟΝΤΟΣ ΤΗΝ ΒΑΣΙΛΕΙΑΝ ΠΑΡΑ ΤΟΥ ΠΑΤΡΟΣ ΚΥΡΙΟΥ ΒΑΣΙΛΕΙΩΝ ΜΕΓΑΛΟΔΟΞΟΥ* 'In the reign of the youth who has inherited the kingship from his father, Lord of Kingdoms great of glory'.<sup>6</sup> 'King' is a common-sense alternant with 'Ptolemy'. Their interchangeability is confirmed by  (line 28 middle), which comprises the demotic equivalent of the long cartouche but with 'king' in place the demotic 'Ptolemy'. The phrase that concludes the demotic passage of the Rosetta Stone is, then, precisely that from line 28 but with a further epithet inserted in the middle.

6. The translation follows Quirke and Andrews 1988.

The absence of the expected demotic correspondent to the long cartouche at the end of the text is therefore not merely not a problem: it is a boon. The interchangeability of  'king' and  'Ptolemy' suggests that the short cartouche is simply 'Ptolemy' alone, without epithets.

#### 4.2. Fixing the Epithets

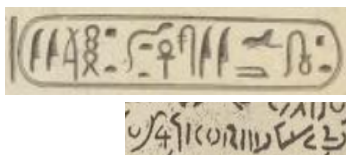


FIGURE 5. Homologies: the long cartouche and its demotic correspondent

With the meaning of the shorter the cartouche possibly fixed, the next question is which epithets embellish 'Ptolemy' in the longer cartouche. Placing the longer cartouche above the corresponding demotic reveals a clear homology in the characters that follow 'Ptolemy' (Fig. 5). The crook of the ankh and the sweep of the serpent are recognisable in both. This is obvious to anyone undertaking a decipherment-level inspection of the material but the underlying principle had previously been identified. Half a century before Åkerblad, Caylus (1752, pp. 70–72) had written:

*les lettres Égyptiennes proprement dites, n'étoient au fond que des hiéroglyphes pareils à ceux des obélisques, mais simplifiés & modifiés par le besoin & par l'usage. ... pour s'en convaincre, on n'a qu'à jeter les yeux sur le N<sup>o</sup>. I. de la XXVI<sup>e</sup>. Planche.*

Egyptian letters, properly so-called, are at root but hieroglyphs like those found on obelisks, yet simplified and modified by need and by usage. ... to convince oneself of this, one need only cast an eye on [Fig. 6].

There is then a mismatch between the hieroglyphs and the corresponding demotic. (As we have already seen with regard to the final cartouche, the two texts are not exact translations.) The demotic is 'Ptolemy'+X and the hieroglyphs, 'Ptolemy'+X+Y. In the Greek, the king's name occurs with fullest epithets as ΠΤΟΛΜΑΙΟΣ ΑΙΩΝΟΒΙΟΣ ΗΓΙΑΠΗΜΕΝΟΣ ΥΠΟ ΤΟΥ ΦΘΑ ΘΕΟΣ ΕΠΙΦΑΝΗΣ ΕΥΧΑΡΙΣΤΟΣ 'Ptolemy, the everliving, beloved of Ptah, the god manifest, the benevolent'. (This counts as three epithets because 'god manifest' and 'benevolent' always cooccur.) Assuming the order of epithets to be invariant between the three Rosetta Stone texts, X could be 'the everliving' and Y, either a second epithet 'beloved of Ptah' or a sequence of two 'beloved of Ptah, the

Pl. XXVI

| Hieroglyphes. | Lettres. | N. 1 <sup>re</sup> | Hieroglyphes. | Lettres. |
|---------------|----------|--------------------|---------------|----------|
| 1.            | ⲙ        | ⲙ                  | 12.           | ⲙ        |
| 2.            | ⲙ        | ⲙ                  | 13.           | ⲙ        |
| 3.            | ⲙ        | ⲙ                  | 14.           | ⲙ        |
| 4.            | ⲙ        | ⲙ                  | 15.           | ⲙ        |
| 5.            | ⲙ        | ⲙ                  | 16.           | ⲙ        |
| 6.            | ⲙ        | ⲙ                  | 17.           | ⲙ        |
| 7.            | ⲙ        | ⲙ                  | 18.           | ⲙ        |
| 8.            | ⲙ        | ⲙ                  | 19.           | ⲙ        |
| 9.            | ⲙ        | ⲙ                  | 20.           | ⲙ        |
| 10.           | ⲙ        | ⲙ                  | 21.           | ⲙ        |
| 11.           | ⲙ        | ⲙ                  | 22.           | ⲙ        |

FIGURE 6. Caylus' hieroglyphic-demotic homologies

god manifest, the benevolent'; or X could be the longer phrase 'the ever-living, beloved of Ptah', in which case Y could only be the remaining epithet 'the god manifest, the benevolent'.

Evidence from Horapollon whittles these options down. His *Hieroglyphica* reports that ⲙ represents 'time',<sup>7</sup> ⲙ 'eternity', and ⲙ 'love' (book 1, chapters 42 and 1 and book 2, chapter 26, respectively; Cory 1840, pp. 5, 64, 104). 'Time' and 'everlasting' support reading X as 'the everliving'. 'Love' supports reading Y as being or including 'beloved of Ptah'.

Scholars from the mid-nineteenth century onwards came to realise that Horapollon was rather unreliable. However, this was not known at Åkerblad's time, before and after which he was a standard source for studies of hieroglyphs (Kircher 1654, Zoëga 1797, Champollion 1824). Åkerblad was not alone in the fealty he afforded Classical sources. He believed, for instance, that demotic had seven vowel characters (Åkerblad, 1802c, p. 56) based on a mere passing illustration of the importance of enunciation in Demetrius' second-century style manual (Roberts, 1902, 104–105, §71):

Ἐν Αἰγύπτῳ δὲ καὶ τοὺς θεοὺς ὑμνοῦσι διὰ τῶν ἑπτὰ φωνηέντων οἱ ἱερεῖς, ἐφεξῆς ἡχοῦντες αὐτά, καὶ ἀντὶ αὐλοῦ καὶ ἀντὶ κιθάρας τῶν γραμμάτων τούτων ὁ ἦχος ἀκούεται ὅπ' εὐφωνίας, ... ἀλλὰ περὶ τούτων μὲν οὐ καιρὸς μηκύνειν ἴσως.

In Egypt the priests, when singing hymns in praise of the gods, employ the seven vowels, which they utter in due succession; and the sound of these

7. A more accurate reading, closer to the actual one of 'life' (ⲙⲏⲕ), would have been available to Åkerblad via the ecclesiastical histories of Socrates Scholasticus (book 5, chapter 17) and Salaminus Sozomen (book 7, chapter 15), which, in relating an early trademark dispute (ankh versus crucifix) between pagans and Christians at the destruction of Temple of Serapis, give the reading 'life to come' (Schaff and Wace, 1890, pp. 127, 386).





In the Greek and its French transcription, the onsets of these names differ: ΠΤΟΛΕΜΑΙΟΣ and *Ptolémée* versus ΦΘΑ and *Phtha*. Åkerblad's native Swedish spellings of these names, resembling the English, might have helped him. But alphabetic vagaries need not detain us. Evidence from Åkerblad's own work shows that he was well able to formulate the hypothesis that the shared hieroglyphs correspond to shared sounds.

An initial starting point might have been Greek grammarians like Aristides Quintilianus, Aristotle, and Dionysius Thrax (Allen, 1968, p. 16). They recorded that, before <φ> and <θ> were sounded as the continuants /f/ and /θ/, they stood for the aspirates /p<sup>h</sup>/ and /t<sup>h</sup>/. The mismatch between 'Ptolemy' and 'Ptah' was not as large as /pt/ to /fθ/ but only /pt/ to /p<sup>h</sup>t<sup>h</sup>/.

The *Lettre* shows that Åkerblad had made this connection and in fact a broader one. Commenting on the demotic spelling of 'Philenos' with an initial *p*, he writes (1802, pp. 24–25):

*La première lettre que nous avons reconnue jusqu'ici pour un P, représente ici le Φ, Ph; ce qui, j'espère, ne souffrira aucune difficulté. Les Égyptiens ont cependant une lettre aspirée qui répond plus particulièrement au Φ des Grecs; mais il paroît qu'ils n'étoient pas très scrupuleux sur le changement des lettres du même organe. Leurs descendans, les Coptes, prennent à chaque instant la même liberté; ils écrivent, par exemple, ΡΕΒΕΡΝΟΒΙ au lieu de ΡΕΦΕΡΝΟΒΙ, &c.*

The first letter, which we have hitherto recognised as a P, represents here a Φ, *Ph*; which, I hope, will cause no difficulty. The Egyptians have, though, an aspirated letter which responds more particularly to the Φ of the Greeks; yet it appears they were none too scrupulous in changing letters of the same organ. Their descendants, the Copts, take the same liberty at every moment; they write, for example, ΡΕΒΕΡΝΟΒΙ [*rebernobi*] instead of ΡΕΦΕΡΝΟΒΙ [*refernobi*], &c.

This passage takes the correlation one step further than is needed here. Åkerblad recognises that letters for homoorganic sounds (by which, most likely, he intends specifically obstruents) are liberally substituted for one another. He cites in support an example of Coptic <B> being used for /f/ and /b/. Elsewhere (Åkerblad 1802c, pp. 48–49; again using Coptic to transcribe the demotic), he addresses the alternation between <π> and <ϙ>, which ordinarily denote /p/ and /f/ respectively, noting that ὠπογορ or ὠφογορ are equally valid representations of 'the king'.

In his response to Åkerblad's *Lettre*, published as an appendix thereto, Silvestre de Sacy wrote (Silvestre de Sacy, 1802b, p. 66):

*Il n'y a assurément rien à dire contre les suppositions par lesquelles vous substituez le Π au Φ, le T au Δ et la K au Γ; et c'est une des idées les plus heureuses que vous ayez pu employer pour vous frayer la voie au déchiffrement de cette inscription.*


There is surely naught to be said against the assumptions by which you substitute Π for Φ [*p* for *f*], T for Δ [*t* for *d*], and K for Γ [*k* for *g*]; and it is one of





finding a second instance of phonography in the hieroglyphs he could tentatively read.<sup>9</sup>

## 5. Plausibility and Impact

In this counterfactual history of decipherment, I have deliberately avoided implausible steps (as in footnote 9), leaps that seem sensible only to modern scholars who know what the ultimate solution was. Extremely able scholars did not take the path sketched above. The next to take up the baton, more than a dozen years after Åkerblad, was Young. His landmark *Encyclopaedia Britannica* article, though substantive in its progress on the script, contained numerous errors, misreading ‘Ptah’ as ‘loved’ for instance (Young, 1824 [1819], pl. 76). Small as this particular mistake is, it would have undone the argument above. It thus cannot be taken as given that Åkerblad would have followed the path we have just traced.

Nonetheless, his doing so is plausible. The steps above are in keeping with his work on the Rosetta Stone demotic and at the same time sympathetic to his general mindset. Åkerblad and Young differed fundamentally in the role they afforded phonography in Egyptian writing. Åkerblad sought to extract the alphabet that he believed had been promised by Plutarch (*Of Isis and Osiris* §56; Thuault 2018). Young regarded phonography as marginal in all Egyptian writing. His encyclopaedia article refers to Åkerblad’s alphabet as ‘supposed’. Later, he was more emphatic: ‘no [explanatory] alphabet would ever be discovered, because it had never been in existence’ (Young, 1823, p. 13). With fewer materials at his disposal, Åkerblad had a more constrained domain to investigate and was more open to phonographic hypotheses alongside the logographic reading his *Lettre* adduced for  (‘first’, ‘second’, ‘third’; section 2). The steps above conform to the reasoning he had employed in his analysis of the demotic and/or build on analysis of his predecessors. Indeed, several crucial insights (especially concerning the interchangeability of letters and crosslinguistic differences of word order) were explicitly formulated by him.

---

9. The correct reading of , namely *mere*, would have supported the same conclusion, given the variable representation of vowels in demotic, which Åkerblad identified, to Silvestre de Sacy’s distaste. However, it is excessively anachronistic to impute this reading to Åkerblad. For *mere*, one must allow for single signs that stand for several sounds at once, like  (*mr* (*mer*)). Although Young advocated such readings, it took evidence from the *Description de l’Égypte* (Commission des sciences et arts d’Égypte, 1809–1822) for him to make the step. Understanding of biliteral signs did not emerge until half a century after Åkerblad (Rougé, 1853).


Had Åkerblad pursued this route, the impact on contemporary understanding would have been substantial, advancing decipherment in five different ways. First, Caylus and Zoëga had argued that cartouches contained names and similar formulae. Åkerblad's reading of the short cartouche as a name and the long one as a name plus epithets would have proven this conjectural argument correct.

The idea that Egyptians would ignore signs' inherent semantics to write names phonetically was not proposed until 1811, in a Chinese-inspired footnote to Silvestre de Sacy's review of the study of Coptic place names that he had favoured over Åkerblad's (Silvestre de Sacy, 1811, p. 184):

*On sait que les Chinois ... sont obligés quelquefois d'employer un certain signe pour avertir que les caractères qui entrent dans l'expression d'un nom propre, sont réduits à cette seule valeur. Je conjecture que dans l'inscription hiéroglyphique de Rosette, on a employé au même usage le trait qui entoure une série d'hiéroglyphes.*


We know that the Chinese ... are obliged at times to use a particular sign to warn that the characters that enter into the expression of a proper name are reduced to this single [phonetic] value. I conjecture that, in the hieroglyphic inscription of Rosetta, the feature that surrounds a series of hieroglyphs was used for the same end.

There are two elements to Silvestre de Sacy's conjecture, one correct, one not.

The correct part is that hieroglyphs used for proper names might be read phonetically. Reading  as 'Ptolemy' only makes sense in phonetic terms. To construe the signs semantically, one would have to claim that a lion on something apparently platform-like, surrounded by reeds and a lasso, themselves flanked by a fold, a square, and a semicircle (now known to be cloth, a mat, and a loaf), represented the meaning 'Ptolemy' to the Egyptian mind. Kircher (1654, 1666) was the *reductio ad absurdum* of such interpretations. By the time of Zoëga, such fantasies had been abandoned but no concrete understanding of how hieroglyphs could be used to write proper names (or much else) had yet been achieved. The phonetic use of hieroglyphs for /p/, /t/, and /ai/ would have been Åkerblad's second contribution, anticipating Silvestre de Sacy's conjecture about cartouches by almost a decade.

Avoiding the incorrect part of Silvestre de Sacy's conjecture would have been Åkerblad's third contribution. He suggested that the cartouche itself signalled to the reader the suspension of semantic reading and the switch to phonography. Champollion (1824) would eventually show that phonetic readings were ubiquitous outside cartouches but Åkerblad could have forestalled the converse half of Silvestre de Sacy's error. The presence of the ankh sign, meaning (from available sources) 'time' or 'life to come', in the phrase 'everliving', would have strongly suggested the semantic readings remained available within cartouches.

The discovery that Egyptians wrote not just foreign names like ‘Ptolemy’ but their native names like ‘Ptah’ phonetically was a pivotal moment in decipherment history, marking Champollion’s decisive break from Young. So important and contentious was it that Champollion did not include it (except as a promissory note) in his 1822 *Lettre*. Instead, it waited until his 1824 *Précis*, a work six times the length of the *Lettre*, where the claim could be elaborated at leisure. Åkerblad’s fourth potential contribution, the discovery of the onset of ‘Ptah’ written phonetically, would have presaged this result by two decades. It might possibly have altered Young’s antiphonetic thinking—in which case, the whole history of decipherment could have looked very different.

Finally, the presence of  read phonetically in ‘beloved’, that is to say, in a word, not a name, might have led to the tentative hypothesis that phonetic writing was not confined to names. This possibility was not mentioned in Champollion’s *Lettre*. It had to await the *Précis* even to be formulated. The conceptual space for this insight could have been earmarked decades earlier by applying Åkerblad’s methods to hieroglyphs.

## 6. Conclusions

Institutions need to guard standards of scholarship. Spurious decipherments—‘the Rosetta Stone demotic is Slavic Macedonian’, ‘the Voynich manuscript is an unknown Romance language’, ‘the Phaistos Disk is early Georgian’ (again, I omit references deliberately)—need to be recognised as such. One way to achieve this is to reexamine past decipherments and the tools that led to them, ensuring that these are accessible, well understood, and subject to repeated verification. I have attempted to do this above by demonstrating that Åkerblad’s methods, which proved successful for demotic, would have yielded material insight into hieroglyphs, advancing the field by some twenty years.

Although guardians of standards, those of us in institutions need also to guard against ourselves, making sure we do not overlook crucial contributions from quarters that we have deliberately or inadvertently excluded. Åkerblad is far from an isolated case in the history of decipherment or academia more broadly. His research foundered not just because his diplomatic obligations took him elsewhere but because he was discouraged and denied opportunities to advance his thought and field.

Decipherment is the bedrock of philography, the study of writing systems. A key philographic finding is the highly convergent evolution of independent writing systems designed for disparate languages. This convergence looks mysterious if we think of writing systems only as the products of specific peoples, languages, cultures, and material resources. Convergence is expected, however, if we view all writing systems as the

product of a single, shared object, the human mind. A proper appreciation of the power of the decipherer's rather spartan toolkit provides important insight into the mental forces that mould writing. So viewed, decipherment constitutes its own pathway to insight into the human capacity for language.

## References

- Adkins, Lesley and Roy Adkins (2000). *The Keys of Egypt: The Race to Read the Hieroglyphs*. London: HarperCollins.
- Agustín y Albanell, Antonio (1587). *Dialogos de medallas, inscripciones y otras antigüedades*. Tarragona: Felipe Mey.
- Åkerblad, Johan David (1802a). *Inscriptionis Phœniciae Oxoniensis nova interpretatio*. Paris: Typographiâ Reipublicæ.
- (1802b). "Lettre de M. Akerblad au C. Silvestre de Sacy, sur la découverte faite par lui de l'écriture cursive copte." In: *Magasin encyclopédique: ou journal des sciences, des lettres et des arts* 7.5, 489–494 plus plate.
- (1802c). *Lettre sur l'inscription Égyptienne de Rosette: adressée au citoyen Silvestre de Sacy, Professeur de langue arabe à l'École spéciale des langues orientales vivantes, &c.* Paris: L'imprimerie de la République.
- (1834 [1810]). "Mémoire: Sur les noms coptes de quelques villes et villages d'Égypte." In: *Journal Asiatique* 13.4–5, pp. 337–377, 385–435.
- Allen, W. Sidney (1968). *Vox Graeca: A Guide to the Pronunciation of Classical Greek*. Cambridge: Cambridge University Press.
- Ameilhon, Hubert-Pascal (1803). *Éclaircissemens sur l'inscription grecque du monument trouvé à Rosette, contenant un décret des prêtres de l'Égypte en l'honneur de Ptolémée Épiphane, le cinquième des rois Ptolémées*. Paris: Institut National.
- Bankes, William John (1821). *Geometrical Elevation of an Obelisk ... From the Island of Philæ, Together with the Pedestal ... First Discovered There by W.J. Bankes ... in 1815: At Whose Suggestion & Expense, Both Have Been Since Removed ... For the Purpose of Being Erected at Kingston Hall in Dorsetshire. [Engravings of the Hieroglyphics and Inscriptions.]* London: John Murray.
- Barthélemy, Jean-Jacques (1759). "Réflexions sur l'alphabet et sur la langue dont on se servoit autrefois à Palmyre." In: *Mémoires de littérature, tirés des registres de l'Académie royale des inscriptions et belles-lettres* 26, for the years 1752–1754, 577–597, pl. I–III.
- (1764). "Réflexions sur quelques monumens phéniciens et sur les alphabets qui en résultent." In: *Mémoires de littérature, tirés des registres de l'Académie royale des inscriptions et belles-lettres* 30, for the years 1758–1760, pp. 405–427.
- Brugsch, Heinrich (1848). *Scriptura Aegyptiorum demotica ex papyris et inscriptionibus explanata*. Berlin: Gærtner.

- Bunsen, Christian Charles Josias (1845). *Ägyptens Stelle in der Weltgeschichte: Geschichtliche Untersuchung in fünf Büchern*. Vol. 1. Hamburg: Friedrich Perthes.
- Caylus, Anne-Claude-Philippe de Tubières-Grimoard de Pestels de Levis (1752). *Recueil d'antiquités égyptiennes, étrusques, grecques et romaines*. Vol. 1. Paris: Desaint et Saillant.
- (1762). *Recueil d'antiquités égyptiennes, étrusques, grecques, romaines et gauloises*. Vol. 5. Paris: Tilliard.
- Champollion, Jean-François (1822). *Lettre à M. Dacier, secrétaire perpétuel de l'Académie royale des inscriptions et belles-lettres, relative à l'alphabet des hiéroglyphes phonétiques employés par les Égyptiens pour inscrire sur leurs monuments les titres, les noms et les surnoms des souverains grecs et romains*. Paris: Firmin Didot Père et Fils.
- (1824). *Précis du système hiéroglyphique des anciens Égyptiens, ou Recherches sur les éléments premiers de cette écriture sacrée, sur leurs diverses combinaisons, et sur les rapports de ce système avec les autres méthodes graphiques égyptiennes*. Paris: Treuttel et Würtz.
- Commission des sciences et arts d'Égypte (1809–1822). *Description de l'Égypte, ou recueil des observations et des recherches qui ont été faites en Égypte pendant l'expédition de l'armée française, publié par les ordres de sa Majesté l'Empereur Napoléon le Grand*. Vol. 1–24. Paris: Imprimerie Impériale.
- Cory, Alexander Turner (1840). *The Hieroglyphics of Horapollo Nilous*. London: William Pickering.
- Desset, François et al. (2022). “The Decipherment of Linear Elamite Writing.” In: *Zeitschrift für Assyriologie* 112.1, pp. 11–60.
- Halley, Edmond (1695). “Some account of the ancient state of the city of Palmyra, with short remarks upon the inscriptions found there.” In: *Philosophical Transactions of the Royal Society of London* 19.218, pp. 160–175.
- Hincks, Edward (1848). “Egypt and the Bible.” In: *The Dublin University Magazine* 32.190, pp. 371–388.
- Jiǎng, Yùbīn 蒋玉斌 (2018). “释甲骨文中的‘蠡’兼论相关问题 [An Elucidation of the Oracle Bone and Bronze Script Character 蠡 and Related Problems].” In: *Fudan Journal (Social Sciences)* 5, pp. 118–138.
- Kircher, Athanasius (1654). *Oedipus Aegyptiacus*. Vol. 3 *Theatrum hieroglyphicum*. Rome: Typographia Vatisil Mascardi.
- (1666). *Obelisci Aegyptiaci: nuper inter Isaei Romani rudera effossi interpretatio hieroglyphica*. Rome: Typographia Varesij.
- Leibniz, Gottfried Wilhelm (1715). “Letter dated 13 January 1714.” In: *Oratio dominica in diversas omnium fere gentium linguas versa et propriis cuiusque linguae characteribus expressa: Una cum dissertationibus nonnullis de linguarum origine variisque ipsarum permutationibus*. Ed. by John Chamberlayne. Amsterdam: Willem & David Goeree, pp. 22–31.
- Leitch, John, ed. (1855). *Miscellaneous Works of the Late Thomas Young*. Vol. 3. London: John Murray.

- Parkinson, Richard (1999). *Cracking Codes: The Rosetta Stone and Decipherment*. with contributions by Whitfield Diffie, Mary Fischer, and R.S. Simpson. Berkeley and Los Angeles CA: University of California Press.
- Quirke, Stephen and Carol Andrews (1988). *The Rosetta Stone: Facsimile Drawing with an Introduction and Translations*. London: British Museum Publications.
- Ray, John (2007). *The Rosetta Stone and the Rebirth of Ancient Egypt*. Cambridge MA: Harvard University Press.
- Roberts, W. Rhys (1902). *Demetrius on Style: The Greek Text of Demetrius de Elecutione Edited after the Paris Manuscript with Introduction, Translation, Facsimiles, etc.* Cambridge: Cambridge University Press.
- Robinson, Andrew (2012). *Cracking the Egyptian Code: The Revolutionary Life of Jean-François Champollion*. Oxford: Oxford University Press.
- Rougé, Emmanuel de (1853). "Mémoire sur l'inscription du tombeau d'Ahmès, chef des Nautoniers." In: *Mémoires présentés par divers savants étrangers à l'Académie* 3, pp. 1–196.
- Schaff, Philip and Henry Wace, eds. (1890). *A Select Library of the Nicene and Post-Nicene Fathers of the Christian Church*. Vol. 2 of series 2, *Socrates, Sozomenus: Church histories*. New York, NY: Christian Literature Company.
- Silvestre de Sacy, Antoine Isaac (1793). *Mémoires sur diverses antiquités de la Perse, et sur les médailles des rois de la dynastie des Sassanides; suivis de l'histoire de cette dynastie, traduite du persan de Mirkbond*. Paris: Le Louvre.
- (1802a). *Lettre au Citoyen Chaptal, Ministre de l'intérieur, Membre de l'Institut national des sciences et arts, &c. au sujet de l'inscription Égyptienne du monument trouvé à Rosette*. Paris: L'imprimerie de la République.
- (1802b). *Réponse du C.<sup>en</sup> Silvestre de Sacy*. appendix to Åkerblad 1802c, 64–70.
- (1811). "Mémoires géographiques et historiques sur l'Égypte et sur quelques contrées voisines, recueillis et extraits des manuscrits coptes, arabes, etc., de la Bibliothèque impériale, par Et. Quatremère, professeur de littérature grecque à l'Académie de Rouen, correspondant de la Société royale de Gottingue, et de l'Institut de Hollande." In: *Magasin encyclopédique: ou journal des sciences, des lettres et des arts* 4, pp. 177–202.
- (1834). *Note préliminaire*. preface to Åkerblad 1834 [1810], 337–338.
- Solé, Robert and Dominique Valbelle, eds. (1999). *La Pierre de Rosette*. Paris: Seuil.
- Thomasson, Fredrik (2013). *The Life of J.D. Åkerblad: Egyptian Decipherment and Orientalism in Revolutionary Times*. Leiden: Brill.
- (2014). "Silvestre de Sacy et J.-D. Åkerblad: Compétition et coopération dans les études égyptiennes." In: *Silvestre de Sacy: Le projet européen d'une science orientaliste*. Ed. by Michel Espagne, Nora Lafi, and Pascale Rabault-F Feuerhahn. Paris: éditions du Cerf, pp. 271–293.



- Thuaault, Simon (2018). "Egyptian Hieroglyphs in Classical Works, Between Pride and Prejudice." In: *Aegyptiaca: Journal of the History of Reception of Ancient Egypt* 3, pp. 191–212.
- Vertue, George et al., eds. (1815). *Vetusta monumenta: quæ ad rerum Britannicarum memoriam conservandam Societas Antiquariorum Londini sumptu suo edenda curavit*. Vol. 4. London: Society of Antiquaries.
- Werning, Daniel A. and Eliese-Sophia Lincke (2019). *The Rosetta Stone Online Project*. <http://hdl.handle.net/21.11101/0000-0001-B537-5>, accessed 26 August 2022.
- Whittaker, Gordon (2021). *Deciphering Aztec Hieroglyphs: A Guide to Nahuatl Writing*. Oakland CA: University of California Press.
- Young, Thomas (1823). *An Account of Some Recent Discoveries in Hieroglyphical Literature, and Egyptian Antiquities: Including the Author's Original Alphabet, as Extended by Mr. Champollion, with a Translation of Five Unpublished Greek and Egyptian Manuscripts*. London: John Murray.
- (1824 [1819]). "Egypt." In: *Supplement to the Fourth, Fifth, and Sixth Editions of the Encyclopædia Britannica*. Ed. by Dugald Stewart et al. Vol. 4. published anonymously. Edinburgh and London: Archibald Constable et al., 38–74 and plates LXXIV–LXXVIII.
- (1830). *Rudiments of an Egyptian Dictionary in the Ancient Enchorial Character; Containing All the Words of Which the Sense has been Ascertained*. Published posthumously as an appendix to Henry Tattam *A compendious grammar of the Egyptian language as contained in the Coptic and Sabidic dialects; with observations on the Bashmuric: together with alphabets and numerals in the hieroglyphic and enchorial characters; and a few explanatory observations*. London: John and Arthur Arch.
- Zoëga, Georg (1797). *De origine et usu obeliscorum ad Pium Sextum, Pontificem Maximum*. Rome: Lazzarini.



# From Clay Tablet to Digital Tablet. The Diamesic Variation of Writing

Sveva Elti di Rodeano

*Abstract.* ‘Diamesy’ is a metalinguistic term referring to the communication. The aim of this paper is to redefine the concept of diamesy applied to writing systems from the grapholinguistic point of view, in order to insert it into the architecture of writing’s variation, and investigating writing shifts in relation to media identity.

For this purpose, the story of the term diamesy, its attestations and meanings in both linguistics and grapholinguistics will be illustrated, providing cases of diamesic variation in writing history. Afterwards, the focus will be on the materiality of writing and the relation between it and material technology, in order to redefine the concept of medium, distinguishing it from mode and modality of communication, and its significance for writing variation.

## 1. Introduction

The concept of diamesy has taken its first steps within the debate about oral and written language, as the identity of the medium used for the communication was found to be a factor responsible for differences in

The term was introduced by Mioni (1983) to emphasize the difference between oral and written modes of representation in contemporary Italian.

In linguistic studies the concept of variation depending upon media has been explored by several critics and reviews, which have the same idea that language can change explicitly and exclusively due to materials, channel, or mode of representation. Therefore the concept of diamesy itself has undergone several and different interpretations (and

---

Sveva Elti di Rodeano  0000-0002-7068-8192

Università Ca’ Foscari Venezia Palazzo Malcanton Marcorà Dorsoduro 3484/d, 30123 Venezia. E-mail: sveva.eltidirodeano@unive.it

Funded by the European Union (ERC-2022-COG, CAnCAn, G.A. 101088363). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 219–236. <https://doi.org/10.36824/2022-graf-elti>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

critiques), which offered reasons to reduce its area of implementation and to adopt other expressions like “medial variation” and “immediacy/distance,” referring to the conditions of communication.

For the purpose of introducing the concept of diamesy as one dimension of grapholinguistic variation and, hence, applying it to writing, it is necessary to review the status of writing in both grapholinguistics and linguistics. Moreover, given that the discussion point about the influence of medium in linguistics was born from the debate about oral and written language, it is also necessary to reassess the status of both writing and speaking in relation to language.

These last issues represent a compulsory and unavoidable topic of research that should be addressed before proceeding. The Greek poet Simonides of Keos, first recorded by Plutarch (*De gloria Atheniensium*, 3.347a), wrote that “Poema pictura loquens, pictura poema silens” (poetry is a speaking picture, painting a silent [mute] poetry). This statement is often rehearsed during investigation into the relation between poetry and painting, which are similar but different artistic products of human imagination. In the same way, it can be stated that speech and writing are similar in their functions, one of which is surely to render language, but they are also different in their nature, one of which is surely the modality of rendering language.

In this paper writing and speaking are considered not to be the same: they work together and serve a similar purpose, which is to convey a message. They are “distinct materializations of language” (Meletis, 2020, p. 72); they do not depend upon each other, because they differ fundamentally, first of all by the fact that writing extends mainly in space, while speech extends in time (Dürscheid, 2016, pp. 24–35).

In this respect, writing is a medium of human communication that involves the representation of a language with written symbols (Ong, 1982). They are meant to render a language into a form that can be reconstructed by other humans separated by time and/or space (Haas, 1996). Hence writing has media for itself, while being a medium for

Above all, this paper recalls, as inspiration, Florian Coulmas’ words: “the media revolution is not just a catchword; it is a reality to which we are forced to adapt and in which writing is of central importance” (2013, p. X). The paper aims to further investigate how medial technologies both constrain and enable writing, and how writing systems, through millennia, have been producing, adapting, and are affected by medial technologies.

## 2. Diamesic Variation in Linguistics

The current classificatory model used in sociolinguistics, especially in European studies, is from the Norwegian linguist Flydal (1952, pp. 241–

258), who introduced the terms *diastraty* and *diatopy*, and the Romanian linguist Coseriu (1955–1956), who, adding the technicism *diaphasy* to the previous terms, elaborated a taxonomy consisting of mutual referring technicisms, forming a structured and cohesive system.

Based on the famous Saussurean dichotomy diachrony/synchrony, Coseriu motivated the prefix and Greek preposition *δια-* “through” in order to give the meaning of internal articulation of the linguistic system, adding Greek and Latin substantives (*τόπος* “place”; *φάσις* “utterance”; *stratum* “a class of society composed of people with similar social, cultural, or economic status”). The resulting scheme, called “architecture of language” had been composed of four linguistic variations: diatopic, diachronic, diastratic, and diaphasic variations.<sup>1</sup> Coseriu coined the only diaphasia technicism, which refers to the different level of formality in communicative situations.

Before the formal introduction of diamesy into the architecture of concept of diaphasic variation into what will then be called diamesic variation. Indeed, Coseriu (1966, p. 199; 1980a, p. 198) discriminated between “language’s style” (first attested in French “styles de langue”), which does refer to the communicative circumstances, and “register,” which should be considered when written/oral/literary language is taken into account. Following his path, several dictionaries have registered distinct headwords for “style” and “register”. Dubois et al. (2002) defined style as “la marque de l’individualité du sujet dans le discours” and as “que ce choix soit conscient et délibéré, ou une simple deviation, le style reside dans l’écart entre la parole individuelle et la langue” (2002, pp. 446–447), and register as “les registres de la parole sont les utilisations que chaque sujet parlant fait des niveaux de langue existant dans l’usage social d’une langue (familier, populaire, soutenu, etc.)”<sup>2</sup> (2002, p. 406).

Likewise, Cardona (1988) defined register as “un determinato livello stilistico (colloquiale, poetico, burocratico, formale e così vi) o un sottocodice relativo ad una lingua speciale”<sup>3</sup> and identified Reid (1956) as the name originator, and style as:

---

1. In German studies *Diatopie*, *Diastratie*, and *Diaphasie* are first attested in Coseriu (1980b, pp. 111–112); in French studies they appeared first in Coseriu (1998, p. 14); in Italian studies the first attestation goes back to 1973 in the Gradi dictionary (cf. Bombi and Orioles (2003, p. 54)).

2. “whether that choice is conscious and deliberate, or a mere deviation, style resides in the gap between individual speech and are the uses that each speaker makes levels of language existing in the social use of a language (colloquial, popular, sustained, etc.)”.

3. “a certain stylistic level (colloquial, poetic, bureaucratic, formal, and so on) or a subcode relating to a special language”.

qualunque manifestazione linguistica, scritta o orale, purché caratterizzata da specifiche scelte (lessicali, sintattiche, eventualmente intonative) all'interno della (...) varietà di riferimento funzionale; si chiamano infatti ss. funzionali (ingl. *functional styles* ecc.) degli insiemi di scelte orientati verso specifici fini comunicativi (s. scientifico, colloquiale, commerciale, ufficiale, giornalistico ecc.).<sup>4</sup>

These dictionaries had registered the subtle but unambiguous distinction once suggested by Coseriu, in order to keep the concepts of informal separate. Indeed, the ultimate aim was to not consider a specific realisation, whether written, oral, transmitted etc., bound to a specific style.

Unfortunately this distinction was not receipted and, hence, “style” and “register” have been often, and still are, treated as synonyms (see § 3.).

For the peculiar Italian linguistic situation, the linguist Mioni (1983) coined and added to this scheme the term *diamesy*, resorting to the Greek μέσος “middle of, between amidst,” with the aim of referring to the expressive medium (written, oral, transmitted etc.) used for the communication.<sup>5</sup> The concept of diamesy has undergone several criticisms, due to the characteristics of the debate that brought the term alongside with the other variation's dimensions: its definition was deeply influenced by the definition of “popular Italian” and “written Italian” (i.e., literary), while no thought was dedicated to the oral opposite poles with no common features at all.

Since its insertion in the variational architecture of language, diamesy has posed considerable semantic and metalinguistic issues, with special reference to its relation with diaphasic variation, as highlighted by Holtus (1984) and Radtke (1992):

Per evitare possibili equivoci è da chiarire che scritto e parlato non vanno intesi come varietà (cioè una deviazione dalla lingua comune), ma come due forme di rappresentazione tramite media diversi (cioè come realizzazioni diverse di una lingua e delle sue varietà)<sup>6</sup> (ibid., p. 67).

4. “any linguistic utterances, written or oral, as long as it is characterized by specific choices (lexical, syntactic, possibly intonative) within the (...) variety of functional reference; in fact they are called functional styles of the sets of choices oriented towards specific communicative purposes (scientific, colloquial, commercial, official, journalistic etc.).”

5. Diamesy is the result of several research projects that, between the 1970s and 1980s, spread in European linguistics. In Italy there has been a debate on so-called “popular Italian” and “regional Italian,” while in France the debate was focused on the diachronic variation of contemporary French (Fusco, 2000).

6. “In order to avoid possible misunderstandings it should be clarified that written and oral must not be understood as varieties (i.e., deviation from common language), but as two forms of representation through different media (i.e., different realizations of a language and its varieties).”

In the same period, Koch and Oesterreicher (1985; 1990) created a different and more articulated model for linguistic variations, which considered the universal parameters of proximity and distance in the communication as the only ones determining the linguistic variation.<sup>7</sup> They were aware of Mioni's works and about diamesy, but they did not agree with the choice of medium as noun formation. Koch wrote:

Der von Mioni (1983, S. 508) eingeführte und in der italienischen und italienistischen Forschung verbreitete Terminus 'diamesisch' ist insofern, wie-wohl aus Gründen der terminologischen Symmetrie recht praktisch, nicht sehr glücklich, weil er auf das Medium (agr. μέσον entsprechend lat. medium) abhebt<sup>8</sup> (ebd., 143, n. 3)

Indeed, they used the term 'medium' meant as "two realizations for linguistic utterances" referring to the phonetic and graphemic realizations, and did discriminate between *Medium* and *Konzeption*, a distinction adopted from the model suited for the French language by Söll (1980), who fixed the general features of oral

Due to their distinction between *Medium* and *Konzeption* and the idea of medium-transferability, the possibility of transferring a communication from one medium to another without any issue (Schneider, 2016; Schneider, Butterworth, and Hahn, 2018), Koch-Oesterreicher's model has undergone several reviews, with special attention to their concept of medium. Indeed, they affirmed that, because language is independent, every text can be transferred in new media without any need of modification. Among the reviewers, Krefeld (2017) has highlighted the weakness of their concept of medium and the ambiguity of it and other terms, such as modality, and observed that the "materialität des Zeichens" must not be confused with "seiner medialität".

On the other hand, Dürscheid (2018) has suggested not using the term 'medium' at all:

Doch vermutlich hätten Koch/Oesterreicher gut daran getan, nicht ihrerseits den Terminus *Medium* zu bemühen; besser hätten sie von Beginn an von *Modalität* gesprochen und folglich von *Modalität und Konzeption*, nicht von *Medium und Konzeption*. Die vielen medientheoretischen Auseinandersetzungen

---

7. Regarding the CMC, Hausendorf, Kesselheim, Kato, and Breitholz (2017, p. 15) have found that this modality of communication goes beyond "face and hear" because it does not necessarily need to be spoken aloud; they then proposed using the terms "presence" (*Anwesenheit*) and "readability" (*Lesbarkeit*), instead of "orality" and "literacy".

8. "The term 'diamesisch', introduced by Mioni (1983, S. 508) and widely used in Italian and Italian research, is not very accurate, although it is quite practical for reasons of terminological symmetry, because it refers to the medium (agr. μέσον corresponding to Latin medium) what? takes off".

rund um ihr Modell wären dann vielleicht ausgeblieben.<sup>9</sup> (Dürscheid, 2018, S. 12)

Moreover, in an attempt to clarify the competing media terms, Dürscheid has introduced three different concepts of medium: medium1, which, refers to the modality, “modalitätbezogen” (ibid., p. 11), constitutes the meant sense found in Koch-Oesterreicher;<sup>10</sup> medium2, which, refers to the technological aspect, “technikbezogen,” meant for the distinction between technologically different media, such as SMS, chat, Internet communication and vocal messages; medium3, which refers to the processing activity for the formation of linguistic signs (Schneider, 2016).

### 3. Diamesic Variation in Grapholinguistics

This terminological uncertainty has led to similar different uses of both diamesy and medium in Grapholinguistics.

First of all, Bunčić, Lippert, and Rabus’ research, edited in 2016, mentioned diamesy with the other dimensions of variations, diastratic and diaphasic. They used this term referring to the Koch-Oesterreicher distinction, while choosing the term “medial,” already used by Dürscheid (2002, pp. 47–50), to refer to the actual distinction in the medium itself.

For the choice of script, in many cases of digraphia the writing material—parchment, wood, stone [...]—plays an important role as well. [...] Such situations can therefore be called medial digraphia (Bunčić, Lippert, and Rabus, 2016, p. 58)

an Italian tradition of referring to a similarly defined kind of variation as diamesic (from Greek μέσος ‘middle’, a cognate of Latin medium). This adjective will be used there to denote a type of digraphia governed by the distinction introduced by Koch and Oesterreicher (1985), viz. diamesic digraphia (Bunčić, Lippert, and Rabus, 2016, p. 59)

In their rich presentation of linguistic cases, regarding diaphasic variation, it is noteworthy to highlight the fact that no distinction is made between style and register (cfr. Bunčić, Lippert, and Rabus (ibid., p. 57)), even if the sense in which the term is used explicitly recalls Coseriu’s interpretation (ibid., n. 25).

---

9. “But presumably Koch/Oesterreicher would have done well not to use the term ‘medium’ themselves; it would have been better if they had spoken of modality from the beginning and consequently of modality and conception, not of medium and conception. The many media-theoretical debates surrounding their model might then have failed to materialize”.

10. This would be the reason for using “modality” and “medium” in the same context and, apparently, with the same meaning.



Afterwards, in his all-embracing grapholinguistics monograph, Meletis (2020) addressed several sociolinguistics issues about writing, mentioning the previous study and the diamesic factor:

Based on the type of opposition—in the Trubezkoyan sense—between two scripts, Bunčić assumes privative and equipollent situations. In (1) digraphia, there is a privative opposition between scripts, meaning one script is lacking a feature that is exhibited by the other script. Which of the two scripts is used in given situations is determined by (1a) diaphasic (pertaining to registers and style), (1b) diastratic (pertaining to social strata), (1c) diamesic (pertaining to the conceptual dimension of written vs. spoken established by Koch and Oesterreicher (1985), or (1d) medial (depending on the writing material) factors. (Meletis, 2020, p. 334)

Then he goes further, recalling the concept of ‘medium’ in Koch and Oesterreicher’s model, that is conceived as distinct from the conceptual dimension.

The hybrid functional nature of both writing and speech is captured by a conceptual distinction that has been impactful in the German-speaking realm: Koch & Oesterreicher’s (1985; 1990; for an English translation, cf. Koch & Oesterreicher 2012) continuum of orality and literacy (cf. also Biber (1988)). In their conception, the dimension of medium—whether a text is medially, i.e., materially, realized in the spoken or written modality—is divorced from the conceptual dimension. (Meletis, 2020, p. 350)

Meletis has explained the reasons lying behind these two terminologies: medial variation refers to the realizations of linguistic utterances,<sup>11</sup> while diamesic variation refers to the modalities and style of the expression.<sup>12</sup> The necessity of such distinctions was already highlighted by

11. Already in Dürscheid (2002, p. 47), referring to Koch-Oesterreicher: “dass eine Äußerung phonisch oder graphisch vorliegt, also gesprochen oder geschrieben wird. In diesem Sinne beziehen sich die Termini ‚mündlich/schriftlich‘ auf “das Medium der Realisierung sprachlicher Äußerungen” (“this simply means the fact that an utterance is phonic or graphic, i.e., it is spoken or written. In this sense, the terms ‘oral/written’ refer to “the medium of realization of linguistic utterances”).

12. In the few lines below, Dürscheid illustrated it: “Zum anderen werde darunter oft der Duktus, die Modalität der Äußerungen verstanden, “kurz: die Konzeption, die die Äußerung prägt” (Koch and Oesterreicher, 1984, p. 587). Es geht dabei um die Tatsache, dass eine bestimmte Ausdrucksweise gewählt wird und diese eher “mündlich” (d.h. an die gesprochene Sprache) oder eher “schriftlich” (an die geschriebene Sprache) angelehnt ist.” (“On the other hand, it is often understood to mean the characteristic style, the modality of the utterance, “in short: the concept that characterizes the utterance” (ibid., p. 587). It is about the fact that a certain mode of expression is chosen and that it is more “oral” (i.e., spoken language) or more “written” (i.e., written language) based”).

Dürscheid: „Zwischen der konzeptionellen und der medialen Dimension von Mündlichkeit und Schriftlichkeit ist also zu unterscheiden“.<sup>13</sup>

Using Coseriu's terminology, in order to define the diphasic variation, the register is here grouped in with the style, which concerns only communicative circumstances. The semantic domain of register, which was supposed to refer to the diamesic variation meant as distinctive for written/oral/literary language, is here combined with the semantic domain of style, which was supposed to indicate situational and functional features. Not accepting the Coseriu's specification, and grouping in style and register for the diaphasic variation leaves no choice other than to create a new term designed for the "conceptual dimension" of written language.

Notwithstanding, this clearcut distinction offers an opportunity to focus on the material features of medium.

Therefore, it can be stated that in Linguistics the actual tendency is either to reabsorb the diamesic variation into the diaphasic one, or to subcategorize it in accordance with the multifaceted concept of medium. Likewise, in Grapholinguistics this has led to the distinction between medial variation, which does refer to the medium intended as material, and diamesic variation, which refers to the intended function, purpose, and conditions of written communication (first prerogative of the style and diaphasic variation).

#### 4. Medium, Mode and Modality

Diamesy is inherently connected to the idea of medium, and its meaning must be reassessed and distinguished from other concepts including modality and mode of transmission.

The concept of medium covers a wide variety of phenomena. It can be seen as the conduit for the transmission of information, and as the form of support for the transmission itself. Ong (1982) objected to a conception of media which reduces them to "pipelines for the transfer of a material called information," because the shape of the pipe affects the type of information that can be transmitted, alters the conditions of reception, and often leads to the creation of works tailor-made for the medium.

In the 20th century, when technological inventions such as photography, film etc. expanded the repertory of channels of communication and means of representation, the concept of medium emerged as an autonomous topic of enquiry, leading to different analysis approaches then called "medium theory". Among the scholars concerned with how the

---

13. "A distinction must therefore be made between the conceptual and the medial dimension of orality and writing".

media altered the meaning of the information transferred through them, McLuhan (1964) stated that media appear to be like an “extension of man,” since they are “forms that shape and reshape our perceptions”. He came to say that “the medium itself is the message”.

Bolter and Grusin (1999) proposed the concept of “remediation” in order to explain relations between different media. In their view, every new technology-based medium must be understood, in the context of new media, as an attempt to “remediate” their limitations and get closer to the elusive goal of “achieving the real”. They did not agree with the claim that every new medium constitutes an improvement over an old one, because every gain in expressions comes at a cost, and new media do not necessarily produce better narratives than older ones.

We have seen that writing itself is a medium of human communication that involves the representation of a language with written symbols. Hence, while being a medium, writing has media for itself, tools and technologies to fulfill its main and first function. It is because writing and technology are so closely linked that technology questions were often overlooked. What is then intended with technology? Technology is not an object, but rather a vital system that is bound to the world of time and space, it is always inextricably tied both to a particular moment in human history and to the practical action of the human-like world in which it is embedded. Is it possible that material technologies, implements, and artifacts can alter and shape the material processes by which writing occurs?

Grapholinguistics should focus on these questions in order to appropriately use the concept of diamesy in writing variation.

#### 4.1. Material Media

Writing is language made material (Haas, 1996, p. 3), hence the relation between writing and material is of high relevance for the definition of writing itself. Writing has its power by linking two powerful systems: the material realms of time and space with the human act of language. Therefore, conceiving writing as inextricably based in the material world can provide a theoretical base from which it is possible to argue about the most recent interaction of the technology question: what is the nature of computer technologies, and what is their impact on writing?

In Grapholinguistics, the concept of writing as medium has been defined due to its nature as realization of language, referring to its materiality.

This type of interpretation highlights the material aspect of medium, which was previously defined with the adjective “medial” and the label “medium2”. Here is Fontanille’s suggestion:

L'extension de l'analyse aux objets-supports et aux situations d'écriture conduit alors à s'intéresser à la structure matérielle du support, à la manière dont elle offre au destinataire une surface d'inscription, et au destinataire, une surface de déchiffrement ou d'action.<sup>14</sup> (Fontanille, 2005, p. 185)

We should go further and consider the writing surface and the writing-bearing object, and, because they are all space- and time-related, the relations between them and the context of storage and display. This interpretation of medium helps to understand how material technologies both constrain and enable writing, and that objects of or with writing are themselves constitutive of meanings, due to the impact of the materiality on human perception.

Looking at modern technological media, such as digital tablets and smartphones, it can be observed that the materiality of the object is nearer to releasing itself from the relation form/function, giving the user an impression of extreme ductility, while the graphic interface looks increasingly like common material media (folder, notebook, paper sheet) and their heavy weights. It has been already observed by Gérard Genette that the material component through which writing can be accomplished offers a "sense supplement" to the text, and that the support's form should be interpreted as one condition for the organization of the text. Genette came to say that "le plus souvent, donc, le paratexte est lui-même un texte: s'il n'est pas encore le texte, il est déjà du texte" (Genette, 1987, p. 9).<sup>15</sup>

The notion that objects are themselves constitutive of meanings, due to the impact of the materiality of their support on perception, can also be discovered in ancient times. For instance, the case of cuneiform script that had been, wherever used and for whatever languages, deeply linked to the clay tablet as bearing object. The tablet in this case has been the common denominator in the spread of cuneiform script and the major medium due to the material, the clay, which was common in these areas, to the easily preservable and portable format, and to the established link between it and bureaucracy. The inscriptions on stone had also played an important role, due to their context of display and intended functions, which was not necessarily to be read, but to express power.

In the cuneiform world there is a strong contrast between the clay tablets, the majority of which come from archive contexts and were probably intended for use by those who could read them, and inscriptions on stone which

14. "The extension of the analysis to support objects and writing situations then leads to an interest in the material structure of the support, the way in which it offers the addressee a surface for inscription, and the addressee, a surface, Decryption, or action".

15. "most often, therefore, the paratext is itself a text: if it is not yet the text, it is already text".

were mostly situated in public or semi-public places and were meant to be seen and to impress a wide range of people including those, probably the majority, who could not actually read them (Matthews, 2013, p. 73)

To use the actual terminology, the medial variation here led to diamesic variation, because it refers to situational features.

Another important aspect of the materiality of writing supports relates to their likelihood of preservation and survival both in ancient times to the present day. Many texts themselves express this trait, for instance the tablet SAA X 373 R. 4-13 (= ABL 334) reads “Let me read the tablets in the presence of the king, my lord, and let me put down on them whatever is agreeable to the king; whatever is not acceptable to the king, I shall remove from them. The tablets I am speaking about are worth preserving until far-off days”.

Moreover, writing tools may also influence a script’s shape. The duc-tus of Indian scripts tends toward straight lines and sharp angles in northern India, for example in Bengali, whereas that of southern Indian scripts, such as Tamil, emphasizes curved lines and rounded forms. The reason is thought to be that the birch bark and paper used in northern India was less prone to being split by a metal stylus drawing straight lines and sharp angles than the palm leaves used in the south. Cuneiform, conversely, has been always written with a stylus, usually obtained from a reed. Its standard name in Akkadian was *qan tuppi* “tablet’s reed”. The way the reed was cut would have determined the calligraphic style. The paleo Babylonian tablets (XVII a.C.) have a typical oblique handwriting, which is due to the use of an oblique cut reed, while Assyrian text was written with a flat cut reed.

## 4.2. Mode of Production and Transmission

Writing is not just a technology (for representing speech) but rather a “mode of communication that is socially learned and culturally shaped or transmitted” (Houston, 2012, p. xiv). Indeed, we have seen that written for communication. Applying this to writing itself implies assuming medium as mode of transmission for writing, and distinguishing medium as technology from medium as communication form, intended as set of social rules that users follow once they have the technologies to use (cf. Meyrowitz (1987)).

This point is relevant for the diachronic perspective of writing variation, because, tracing back through the evolution of written signs, we notice that they have been under the influence of several factors, some due to their physical form, to the physical form of the carrying objects, and to the physical form of the tool implied. This interpretation of medium can then point out the graphetic features concerning material

aspects of medium which influence the writing process. Indeed, signs' shapes, once they are recognized in their signified meanings, become increasingly subject to forces related to movement and perception that change characters written by hand.

These forces include the so-called "biomechanics of production" which Overmann includes in the forces related to movement and perception that change characters written by hand.

In a literate brain, the region with an evolutionarily provided function for recognizing physical objects becomes trained to recognize written characters as if they were physical objects, interpret them through the gestures of hand-writing, and associate them with the meanings and sounds of language. Such reorganization involves not just brains but behaviors and material forms as well.

Biomechanics of production: the use of hands and arms, as well as head and body positions that affect how objects used for writing are held, oriented, viewed, and manipulated (Overmann, 2021, p. 98)

For instance, in proto-cuneiform script, namely the archaic signs attested in Uruk IV (3500-3300 BC) and Uruk III (3300-3000 BC), two general tendencies can be observed: first, the pictographic signs of Uruk IVa become increasingly abstract in Uruk III, as the round and incised strokes are replaced with straight and impressed lines; second, lines' orientations are chosen instead of others, because they allow a more natural flow of the cuneus and require less effort to the scribe. The motivation of both graphemic variations is to minimize the effort to produce writing and make it more efficient. These ultimately link to the nature, material and functional, of the medium.

It was thus the more efficient use of tools that forced the elements that make up signs into their wedgelike shapes, taking on the characteristic angular form. Cuneiform writing therefore originated because of the difficulties of representing curved lines on the fresh clay and the need to break up the signs into segments made up of small rectilinear incisions with a triangular head.

### 4.3. Modality of Writing, Modality of Language

Lastly, writing has been also as a modality of language, written modality of rendering writing, i.e., written language. The modality is "the particular physical means by which an alphabet is executed and received" (Watt, 1983, p. 1543). It is related both to the process of coding and decoding the message and to the intended audience and recipient.

Going back to the cuneiform example, we have seen that physical constraints, due to the support and the tool, are key factors for the graphemic change of script. The law of least effort, or Zipf law, indeed

points to the same direction. Notwithstanding this, for writing there are more features to be considered.

In the case of Egyptian scripts, for instance, the time investment was also an important concern, given the development of cursive scripts rather than Hieroglyphs. The purpose of these was to make coding easier and quicker than what was possible with hieroglyphs. However, writing speed become inversely proportional to legibility, a factor that is directly related to the intended readership: the larger intended readership, the more easily readable the script has to be. Indeed, after demotic was introduced, scribes had to learn it daily and before others, as Clement of Alexandria shows us.

Αὐτίκα οἱ παρ' Αἰγυπτίοις παιδευόμενοι πρῶτον μὲν πάντων τὴν Αἰγυπτίων γραμμάτων μέθοδον ἐκμανθάνουσι, τὴν ἐπιστολογραφικὴν καλουμένην· δευτέραν δὲ τὴν ἱερατικὴν, ἣ χρῶνται οἱ ἱερογραμματεῖς· ὑστάτην δὲ καὶ τελευταίαν τὴν ἱερογλυφικὴν [...]. (*Stromata* V, iv, 20-21)<sup>16</sup>

Therefore, in the later periods of Egyptian writing history, scribes had to take into consideration the intended readership to determine the reading capability and then to choose which script use.

A clear example in diachronic variation is the Ptolemaic sacerdotal decrees, which are engraved on stone in hieroglyphs, demotic, and Greek. In particular, the Decree of Canopus (238 BC) refers to Hieroglyphs as “the script of the pr-‘nh” (hieroglyphic «sh} n pr-‘nh», demotic «sh} (n) pr-‘nh», ἱερός in Greek); to Demotic as “the document script” (hieroglyphic «sh} n š‘.t», demotic «sh} (n) š‘(t)», Greek Αἰγύπτιος); and to Greek as “the script of the Aegean islanders” (i.e., Greeks) (hieroglyphic «sh} n h}w-nb.wt», demotic «sh} (n) wynn», Greek ἑλληνικοῖς (sc. γράμμασιν). The decree of Memphis (196 BC) refers to Hieroglyphs as “the script of the divine words” (hieroglyphic «sh} mdw-nṯr»; demotic «sh} md(.t)-nṯr», Greek: ἱερός); to Demotic as “the document script” (hieroglyphic «sh} n š‘y», demotic «sh} (n) š‘.t», Greek: ἐγχώριος); and to Greek as “the script of the Aegean islanders (i.e., Greeks)”. Both of these documents refer to Demotic as document script, translated in Greek as ἐγχώριος “indigenous.” The Demotic script records are then intended to record everyday business and to be separate to the other two scripts by means of distinct functionality. This means that the nature of medium, conceiving its material constraints, influenced the diachronic evolution of Egyptian scripts, alongside the purpose of the same script, which has changed depending upon the intended subject of the written communication and its intended readership.

16. “Now those instructed among the Egyptians learned first of all that style of the Egyptian letters which is called Epistolographic; and second, the Hieratic, which the sacred scribes practice. And last of all, the Hieroglyphic [...]”.

The increasing use of emojis in digital writing, thanks to the inclusion in the Unicode Standard in 2010 (cfr. Dürscheid and Meletis 2019), is challenging the principle of least effort regarding the time and exertion needed for the production of written signs, and is enlarging the variants of written digital communication. Nowadays the relationship image/writing and the same concept of writing are evolving thanks to the number of emojis included in the Unicode Standard, which will definitely change the intended readership or the capability of read digital written communication.

## 5. Conclusion

Overall, retracing the history of the term *diamesy*, as often happens with terminology, gives us the chance to examine in depth another term, in this case ‘medium’. We have seen that, from a linguistic point of view, the same writing has been seen as medium for language, the influence on which is still undergoing several interpretations. From a grapholinguistic point of view, on the other hand, the concept of medium has been conceived based on the influence that it could have on written functional and situational features.

Now, we have seen that these aspects are inherently bound to each other, given the spatial and timing-related nature of media, and because material and technological aspects inevitably lead to functions which can either directly affect the writing or script choice, or have indexical meaning that affects the communication, then dealing with the temporal and spatial distance or proximity between the different participants in the act of communication.

We are aware that writing has been defined as a technology that extends human ability to communicate with others across space and through time (Haas, 2013). Writing turns the time of communication—the one required by a vocal message, for instance—into space—the one required by a text message—or, I might better suggest, writing adds the time of space to send and receive a message to the space of time that coding and decoding a message need.

The two dimensions of writing, time and space,<sup>17</sup> are possible thanks to the medium, that exists in space, because it is material—whether it is tangible in the common and direct sense, as a paper sheet, or less, as the size in gigabyte of text file—and time, because it is supposed to last over time—for clay tablets over centuries, for digital tablets too, even if the latter is not designed to make texts last a great amount of time.

---

17. Innis (1951) has argued that most media of communication have either a “time bias” or a “space bias”: they have a tendency either toward lasting a long time or toward moving across space.



Moreover, the surface comes first, whether it is something arbitrarily chosen from existing things or a created artifact, it is a space that had to be invented, accepted, and integrated within society. Today becoming literate is still a matter of interacting with material forms: typing on keyboards instead of handwriting with tools might affect motor skills including hand/eye coordination and signs' recognition, because the potential loss of tolerance for ambiguity in how signs are formed will lead to the increasing difficulty of reading handwriting text.

Ultimately, the increasing inclusion of electronic media will involve changes both in individual components of literacy, namely the material forms used for reading and writing, and all interpersonal communicative systems. Because literacy took centuries to develop, it remains an open question how far and deeply electronic media will change it; Florian Coulmas' words can be stated for sure, that "the electronic media revolution has changed and continues to change the linguistic landscape and the public sphere. Written language is at the center of this revolution" (Coulmas, 2013, p. 38).

For our purpose, we might be willing to thank new media and writing tools, which always instigate linguistic innovation beyond the incessant pace of language change (*ibid.*, p. 128) and motivate metalinguistic reflections that lead us now to say that the features constitutive of the same concept 'medium' go beyond the mere materiality of it, and that this should be included in the connotation of the term 'diamesy', which, etymologically, meant everything that stays in between, in the middle of what writing is intended to accomplish.

## References

- Biber, Douglas (1988). *Variation across speech and writing*. Cambridge: CUP.
- Bolter, J. D. and R. Grusin (1999). *Remediation. Understanding New Media*. Cambridge, MA: MIT Press.
- Bombi, R. and V. Orioles (2003). "Aspetti del metalinguaggio di Eugenio Coseriu. Fortuna e recepimento nel panorama linguistico italiano." In: *Studi in memoria di Eugenio Coseriu*. Ed. by V. Orioles. Vol. 10. Plurilinguismo contatti di lingue e culture, pp. 53–73.
- Bunčić, D., S. L. Lippert, and A. Rabus, eds. (2016). *Biscriptality, a sociolinguistic typology*. Heidelberg: Winter.
- Cardona, G. R. (1988). *Dizionario di Linguistica*. Armando Editore: Roma.
- Coseriu, E. (1955–1956). "Determinación y entorno. Dos problemas de una lingüística del hablar." In: *Romanistisches Jahrbuch* 7, pp. 29–54.
- (1966). "Structure lexicale et enseignement du vocabulaire." In: *Actes du premier Colloque international de Linguistique appliquée. Organisé par la Faculté des Lettres et des Sciences humaines de l'Université de Nancy*, pp. 175–217.

- Coseriu, E. (1980a). "'Historische Sprache' und 'Dialekt'." In: *Dialekt und Dialektologie. Ergebnisse des internationalen Symposiums "Zur Theorie des Dialekts," Marburg/Lahn, 5.-10. September 1977*. Ed. by J. Göschel, P. Iviæ, and K. Kehr. Wiesbaden, pp. 106–122.
- (1980b). "Interdisciplinarietà e linguaggio." In: *L'accostamento interdisciplinare nello studio del linguaggio*. Milano: Franco Angeli, pp. 43–65.
- (1998). "Éditorial. Le double problème des unités "dia-s". In: *Les Cahiers dia. Études sur la diachronie et la variation linguistique* 1, pp. 9–16.
- Coulmas, F. (2013). *Writing and Society. An Introduction*. Cambridge: Cambridge University Press.
- Dubois, J. et al. (2002). *Dictionnaire de linguistique*. Paris: Larousse.
- Dürscheid, C. (2002). *Einführung in die Schriftlinguistik*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- (2016). *Einführung in die Schriftlinguistik*. 5th ed. Vol. 5. Göttingen: Vandenhoeck & Ruprecht.
- (2018). "Koch/Oesterreicher und die (neuen) Medien—Anmerkungen aus germanistischer Sicht." In: ed. by T. Gruber et al., pp. 60–81.
- Dürscheid, C. and D. Meletis (2019). "Emojis. A Grapholinguistic Approach." In: *Graphemics in the 21st Century. Brest, June 2018. Proceedings*. Ed. by Y. Haralambous. Brest: Fluxus Editions, pp. 167–183.
- Flydal, L. (1952). "Remarques sur certains rapports entre le style et l'état de langue." In: *Norsk Tidsskrift for Sprogvidenskap* 16, pp. 241–258.
- Fontanille, J. (2005). "Du support matériel au support formel." In: *L'écriture entre support et surface*. Ed. by I. Klock, J. Fontanille, and M. Arabyan. Paris: L'Harmattan, pp. 183–200.
- Fusco, F. (2000). "Français avancé, français populaire, français branché. Varietà e variabilità nel francese contemporaneo." In: *Plurilinguismo* 7, pp. 63–82.
- Genette, G. (1987). *Seuils*. Paris: Seuil.
- Haas, C. (1996). *Writing Technology. Studies on the Materiality of Literacy*. Routledge: New York/London.
- (2013). *Writing technology. Studies on the materiality of literacy*. Routledge: New York.
- Harper, R. F. and L. Waterman (1892–1914). *Assyrian and Babylonian letters belonging to the Kouyunjik Collection of the British Museum*. 14 vols. Chicago: University of Chicago Press.
- Hausendorf, H. et al. (2017). "Textkommunikation. Ein textlinguistischer Neuansatz zur Theorie und Empirie der Kommunikation mit und durch Schrift." In: *Reihe Germanistische Linguistik* 308.
- Holtus, G. (1984). "Codice parlato e codice scritto." In: *Il dialetto dall'oralità alla scrittura. Atti del XIII Convegno per gli Studi Dialettali Italiani (Catania-Nicosia, 28 settembre 1981)*. Pisa: Pacini, pp. 1–12.

- Houston, D. (2012). *The shape of script. How and why writing systems change*. School for Advanced Research Press: Santa Fe.
- Innis, Harold (1951). *The Bias of Communication*. Toronto: Toronto University Press.
- Koch, P. and W. Oesterreicher (1984). "Schriftlichkeit und Sprache." In: *Schrift und Schriftlichkeit. Ein interdisziplinäres Handbuch internationaler Forschung. An Interdisciplinary Handbook of International Research (ITALIC)*. Ed. by H. Günther and O. Ludwig. Berlin/New York: de Gruyter, pp. 587–604.
- (1985). "Sprache der Nähe—Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte." In: *Romanistisches Jahrbuch* 36, pp. 15–43.
- (1990). *Gesprochene Sprache in der Romania. Französisch, Italienisch, Spanisch*. Vol. 31. Romanistische Arbeitshefte. Berlin, New York: De Gruyter.
- (2012). "Language of Immediacy—Language of Distance: Orality and Literacy from the Perspective of Language Theory and Linguistic History." In: *Communicative spaces: Variation, contact, and change—Papers in honour of Ursula Schaefer*. Ed. by Claudia Lange, Beatrix Weber, and Göran Wolf. Frankfurt a. M.: Peter Lang, pp. 441–473.
- Krefeld, T. (2017). "Rezension zu. Feilke, Helmuth & Mathilde Hennig (Hrsg.). Zur Karriere von ‚Nähe und Distanz‘. Rezeption und Diskussion des Koch-Oesterreicher-Modells." In: *Zeitschrift für Rezensionen zur germanistischen Sprachwissenschaft*. Germanistische Linguistik 10.
- Matthews, R. (2013). "Writing (and Reading) as Material Practice. The world of cuneiform culture as an arena for investigation." In: *Writing as Material Practice. Substance, surface and medium*. Ed. by K. E. Piquette and R. D. Whitehouse. London: Ubiquity Press, pp. 65–74.
- McLuhan, M. (1964). *The Medium is the Message. An Inventory of Effects*. Bantam Books: Toronto.
- Meletis, D. (2020). *The Nature of Writing. A Theory of Grapholinguistics*. Vol. 3. Grapholinguistics and Its Applications. Brest: Fluxus Editions.
- Meyrowitz, J. (1987). *No sense of place. The impact of electronic media on social behavior*. New York: Oxford University Press.
- Mioni, A. (1983). *Italiano tendenziale. Osservazioni su alcuni aspetti della standardizzazione*. Pisa: Pacini, pp. 495–517.
- Ong, Walter J. (1982). *Orality and Literacy. The Technologizing of the Word*. London: Routledge.
- Overmann, K. A. (2021). "Writing system transmission and change. A neurofunctional perspective." In: *Signs—Sounds—Semantics. Nature and transformation of writing systems in the ancient Near East. Papers presented at the 64th Rencontre Assyriologique Internationale, University of Innsbruck, July 2018*. Ed. by G. Gabriel, K. A. Overmann, and A. Payne. Vol. 13. Wiener Offene Orientalistik. Vienna: Ugarit, pp. 99–116.

- Parpola, Simo, ed. (1987). *State Archives of Assyria*. Helsinki: Helsinki University Press.
- Radtke, E. (1992). "Varietà dell'italiano." In: *La linguistica italiana degli anni 1976-1986*. Ed. by A. Mioni and M. A. Cortelazzo. Roma: Bulzoni, pp. 59-74.
- Reid, T. B. W. (1956). "Linguistics, Structuralism and Philology." In: *Archivum Linguisticum* 8.1, pp. 28-37.
- Schneider, J. G. (2016). "Nähe, Distanz und Medientheorie." In: *Zur Karriere von „Nähe und Distanz“. Rezeption und Diskussion des Koch-Österreichers Modells*. Ed. by H. Feilke and M. Hennig. Vol. 306. Reihe Germanistische Linguistik. Berlin: De Gruyter, pp. 333-356.
- Schneider, J. G., J. Butterworth, and N. Hahn (2018). *Gesprochener Standard in syntaktischer Perspektive. Theoretische Grundlagen—Empirie—didaktische Konsequenzen*. Vol. 99. Stauffenburg Linguistik. Tübingen: Stauffenburg.
- Söll, L. (1980). *Gesprochenes und Geschriebenes Französisch*. Grundlagen der Romanistik. Berlin: Erich Schmidt.
- Watt, W. C. (1983). "Mode, modality, and iconic evolution." In: *Semiotics Unfolding. Proceedings of the Second Congress of the International Association for Semiotic Studies, Vienna, July 1979*. Ed. by T. Borbé. Vol. 68. Approaches to Semiotics. Berlin/Boston: De Gruyter, pp. 1543-1550.

# The application of grapholinguistics in palaeography. A case study: Croatian Glagolitic and Cyrillic palaeography


Kristian Paskojević

*Abstract.* Palaeographic research based on grapholinguistics is a relatively new approach within the framework of Croatian palaeography. To accept writing as a form of expression equal to speech, it was necessary to redefine it and create an adjusted description of it. In the forefront of observation, *script* as a complete material ready to be subjected to structural analysis is replaced with *writing*, a cognitive process that transmits a message from an author to a reader/listener. This view on palaeographic research required a different methodological approach, which was successfully resolved with the invention of *palaeographic categories*. These categories are not an entirely new invention; by observing research material through them, the process of script development became clearer. This method has already been tested on numerous documents and charters from the Croatian Mediaeval period. The material analysed in this research was written in either Glagolitic or Cyrillic script. The main goal of this paper is to present current research related to Croatian Mediaeval literacy that uses this methodology.

## 1. Introduction

As scientific disciplines, palaeography and grapholinguistics have a great deal in common. If we examine the definition of grapholinguistics as “the linguistic sub-discipline dealing with the scientific study of all aspects of written language” (Neef 2015: 711), it becomes apparent that the main connection between them is their shared field of research—writing systems. This particular field of interest is probably one of the main reasons they have been marginalised by “mainstream” scientific fields such as linguistics and history. To some extent, these disciplines are still marginalised and viewed as “auxiliary” sciences. Also, both palaeography and grapholinguistics have a marked interdisciplinary character. Historically, palaeographic research has been conducted

---

Kristian Paskojević  0000-0001-7953-0511  
Staroslavenski institut (Old Church Slavonic Institute), Demetrova 11, 10 000 Zagreb,  
Croatia  
E-mail: kristian.paskojevic@stin.hr

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 237–252. <https://doi.org/10.36824/2022-graf-pask>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

for more than three hundred years (taking Mabillon's "De re diplomatica libri VI" [1681.] as the starting point). Throughout this lengthy time period, palaeography developed its own terminology and research methods; with the newest technological advances, it launched itself into the sphere of the *digital humanities*.<sup>1</sup> The grapholinguistic method in Croatian palaeography is a relatively new approach. The key work responsible for its introduction, which correctly places Croatian literacy in the broader European context, is Mateo Žagar's *Grafolingvistika srednjovjekovnih tekstova* (Grapholinguistic of Mediaeval texts, 2007). This book also serves as a solid starting point and methodological reference for grapholinguistic-based palaeographic research. Žagar is known as one of the most productive authors in the field of Croatian Slavic Palaeography, and has written numerous grapholinguistic studies of Glagolitic and Cyrillic written monuments, most of which hail from the Middle Ages.

## 2. The Methodology of Grapholinguistic-Based Palaeographic Research

Palaeographic research conducted in this way uses various palaeographic categories as its main methodological tool. These categories are: letter coordination in the linear system and the general characteristics of a given script; special letter forms; word dividers (punctuation and capitalization, use of blank space, and separation of words in texts); ligatures and abbreviations; the writing of numbers in texts. These have been established as the parameters of systematic palaeographic description, and as such, they play a very important role. The category of *coordination in the linear system* primarily implies the process of simplifying and aligning the letter lines in the central part of the system and the development of letter forms within this system with the aim of achieving optimal writing speed while maintaining recognizable, easily read letter shapes. The phenomenon of coordination is necessary for the writing to literally "flow" as quickly and as efficiently as possible. Coordination differs according to letter type, and in some cases, the nature of the document plays a key role in the creation of this process (e.g., coordination in liturgical texts differs greatly from that in diplomatic charters).

The category of *special letter forms* is mainly focused on the morphology of the letter itself. Research on the morphological characteristics

---

1. The digital humanities (DH) is an area of scholarly activity at the intersection of computing or digital technologies and the disciplines of the humanities. It includes the systematic use of digital resources in the humanities, as well as the analysis of their application (Drucker 2013: 9).

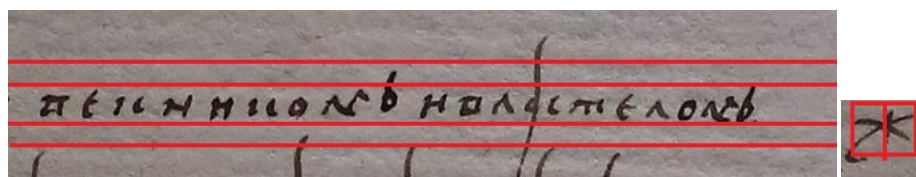


FIGURE 1. Example of a four-line system and its letter module, *Transcript of Emperor Stefan Dušan's Charter* (1352.), Ragusan scribe Đivo Parmezan

of letter forms in various documents allows us to better identify different chanceries and writers, sometimes even different timespans within which a certain letter was used. This is the most important category in grapholinguistic-based palaeographic research.

The separation of individual words in texts with blank spaces that serve as boundary indicators, together with other word dividers, is one of the main conditions for a writing system to be successfully understood. The aim of the study of *word dividers: punctuation and capitalization, use of blank space, and separation of words in texts* is to identify the process by which separate writing was established in a script. The use of word dividers and capital letters is incorporated into every modern European orthography, and their use strongly reflects the civilizational achievement of a given script. Their purpose is to visually optimize a written language message, enabling readers to easily parse the text, both in the case of voiced and silent reading.

Abbreviations, whether contractions, superscriptions, or suspensions, are the result of a writing process derived from the writer's intention to save space on the page. These are largely the result of the ideological motivation that the "sacred words" (sacrament) should not be written in their entirety, following the Jewish principle of not pronouncing the Lord's name. Ligatures also belong to the category of abbreviations, but represent a somewhat different phenomenon. Unlike other abbreviations, ligatures are always generated by a combination of two letters, creating a new, specific letter form.

The *category of writing numbers* identifies all the methods used to write numbers (be it Roman numerals, Arabic numerals, or letters in some scripts like Cyrillic and Glagolitic) and any specifics regarding their writing (if they exist). The provided data is a helpful tool by which to draw parallels between writing methods and various writers and chanceries. This can, of course, also help in the dating of documents—for example, the increased use of Arabic numerals in Croatian history began in the 15th century.<sup>2</sup>

2. Although there are some examples from the late 13 ct. in documents originating from Dubrovnik's chancery (Novak: 293.).

### 3. About the Origins of the Glagolitic Script

Glagolitic script is a unique phenomenon in the world of palaeo-Slavic studies. It shows no immediately apparent similarity to any other known script, and the details of its origins have not yet been fully ascertained. The issue of its creation is one of the most complex questions in the entire field of palaeo-Slavic studies. The main creation theories are divided into three categories—exogenous, endogenous, and a combination of the two (exogenous-endogenous). Exogenous theories are based on the attempt to prove that another script known in the 9th-century Byzantine empire was used as the foundation of Glagolitic script. The most common theories were the Western or Latin theory (including the “St. Jerome theory”), the Gothic or migration theory, the Syrian theory, the Georgian theory, the Armenian theory, etc. The theory claiming that the root of Glagolitic script lies in 8th- and 9th-century Greek minuscule has long been the most widely accepted. This theory was first put forth by Isaac Taylor and Vatroslav Jagić (hence the name *Taylor-Jagić theory*). It survived until the 1970s, when the endogenous view came to the fore (Žagar 2021: 79, 106). The endogenous theories assumed an original, individual approach to the creation of Glagolitic script and the creation of a universal principle according to which its letters were developed, composed, and combined. The combination of three Christian symbols—the circle, the triangle, and the cross—is the basis of Georg Tschernochvostoff’s theological idea of the origins of Glagolitic script. The idea that all the characters in Glagolitic script share a unique graphetic schema (module) was first introduced in 1982 by Vasil and Olga Yonchev. Their attempt to reconstruct a unified letter module from which all Glagolitic letters were derived also included a search for symbolism, but not to the same extent as Tschernochvostoff’s. The proposed wheel/rosetta-shaped module consists of a combination of the symbols of the cross (the letter *a*), Saint Andrew’s cross (X), and the circle. It is also necessary to emphasize that Yonchev correctly recognized that the oldest Glagolitic script was in essence an uncial, two-line script. One of the most important discussions on the origins and development of Glagolitic script was written by Austrian palaeographer Thorvi Eckhardt (1955). She opposed the Taylor-Jagić theory and shifted her research focus to the creative world of Saint Constantine (Cyril), who was most likely the author of Glagolitic script. She focused on the process of writing, discarding the static observation of individual letters.<sup>3</sup>

Of all these theories, the most plausible today is that Glagolitic script was authored by St. Cyril (Constantine), who created it with under

---

3. This idea is crucial for the development of grapholinguistic-based palaeographic approach.



the influence of some contact scripts (such as Greek minuscule script, Georgian, Armenian, etc.). This claim provides the best compromise between the exogenous and endogenous theories by connecting them. As St. Cyril was a well-educated scholar, he likely knew many scripts, the letters from some of which were surely an inspiration in the creation process of Glagolitic script. Yonchev's theory and his creation of a geometric letter module should not be dismissed. However, bearing in mind the broad distribution of this geometric element (wheel/rosseta), it is difficult to say if it was indeed the author's intent to create letters according to this module or if the reconstruction of this schema was a byproduct of a deductive research process.

#### 4. The Development of Angular Glagolitic Script and Grapholinguistic-Based Research in Croatia

Within the framework of Croatian Glagolitic palaeography, the most exciting event to occur between the late 12th and early 13th century was the creation of the angular version of Glagolitic script. Unlike the earlier, rounded Glagolitic script, the angular version was almost exclusively used along the modern-day Croatian coastline<sup>4</sup> and part of the hinterland situated west of the river Krka, while Cyrillic script became dominant east of this border. This is the most simplified explanation of the Glagolitic/Cyrillic diachronic aspect of Croatian Mediaeval literacy. Latin script is the third part of this equation, which was best defined by Eduard Hercigonja as "the triliterate and trilingual culture of the Croatian Middle Ages" in his book of the same title (2006).

The development process of angular Glagolitic script is much easier to understand if observed through the category of letter coordination in the linear system. The formation of the letter module in any script is directly related to its linear organisation. In the case of rounded Glagolitic script, the two-line linear organization was responsible for the coordination of the letters and the creation of the recognizable letter module presented by Yonchev. In 13th-century European literacy, the script was *minusculed*. Because of the rise of scribal awareness that written messages should be delivered faster, advanced writing techniques and

---

4. There were two unsuccessful attempts at introducing Glagolitic script to the western Slavs. The first was that of Charles IV of Bohemia (1346-1378), who, having received the Pope's permission to reintroduce Slavic liturgy, invited a group of Benedictine monks from Glagolitic parishes in Dalmatia to the Emmaus monastery in Prague. This monastery was later destroyed, together with the initiative, during the Hussite wars. The second was that of King Casimir the Great of Poland (1333-1370), who invited Glagolitic monks to the Benedictine monastery in Kleparz near Kraków (Schenker 1995: 165-166).

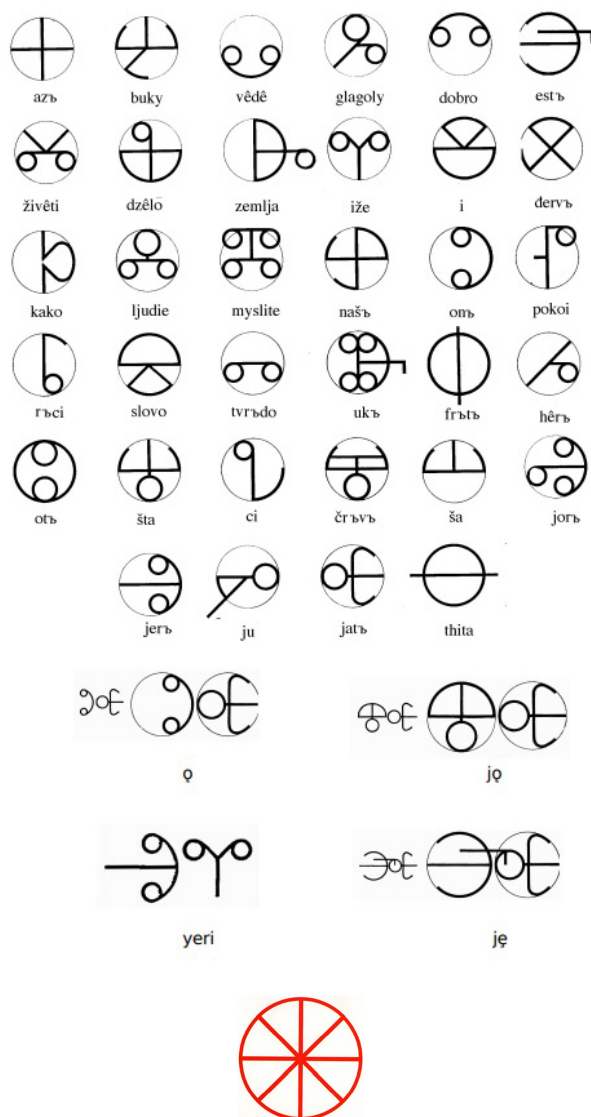


FIGURE 2. A basic portrayal of the rosetta-shaped Glagolitic letter module and the ideal letter forms of round Glagolitic script according to Vasil and Olga Yonchev (1982)

the development of angular cuts to quill tips conditioned a transition to a four-line system of linear text organization, in which the original two lines became the central part of the new system. The main letter parts are coordinated in this central area, while the so-called weak parts (mostly hyper-extended letter lines that are not crucial to the recognition of the letter itself, but are helpful to visual orientation in the text) breach the upper or the lower line (ascenders/descenders), sometimes even both.

The adjustment of complex angular Glagolitic letters to the four-line system created a new square letter module divided into six equal parts. This module was not an exact graphic orientation; it emerged as a concept developed through the minusculation of letter forms. This process also initiated the transition from rounded to angular letter lines (after which these script types are named), the most significant morphological change to take place in the script's history. With it came a change in the letter fields, which opened the way for the more extensive use of ligatures (a special form of abbreviation wherein a combination of two or more letters results in a new letter form). Palaeographic research, such as that conducted under the Scientific Center of Excellence for Croatian Glagolitism, shows the great frequency of ligature writing in Glagolitic liturgical books. For example, The Second Beram Breviary has 322 ligature combinations, which were recorded a total of more than 30,000 times across its 264 folios. This meticulous palaeographic research on the First and Second Beram Breviary (the study on the First Beram Breviary has yet to be printed) is a representative example of recent study of Croatian Glagolitic literacy that is completely based on the grapholinguistic approach. Soon, the Scientific Centre of Excellence for Croatian Glagolitism plans to complete two similar research projects on the First and Second Beram Missal, thus presenting the literary wealth of the Beram Scriptorium.<sup>5</sup> These projects are, by their nature, not merely palaeographic. Each edition includes a printed facsimile and transliteration, studies of morphology, syntax, writing systems, phonology, and vocabulary. In addition to the printed editions, digital databases contain a virtual dictionary with a grammatical and morphological description of each word.<sup>6</sup>

Recent research has placed a greater focus on the graphetic organisation of Glagolitic texts (the visual placement of text on the page, including the division of the text into lines, the formation of word blocks/abandonment of *scriptura continua*, abbreviations, punctuation, the inclusion of different letter sizes and types, the use of blank space as

---

5. Beram is a hill settlement in the central part of the Istrian peninsula. It was one of the most important places in Croatian Medieval Glagolitic literacy.

6. Transliterations are available at <https://beram.stin.hr/>. Public access to the dictionary and other data is still unavailable due to the site being under construction.

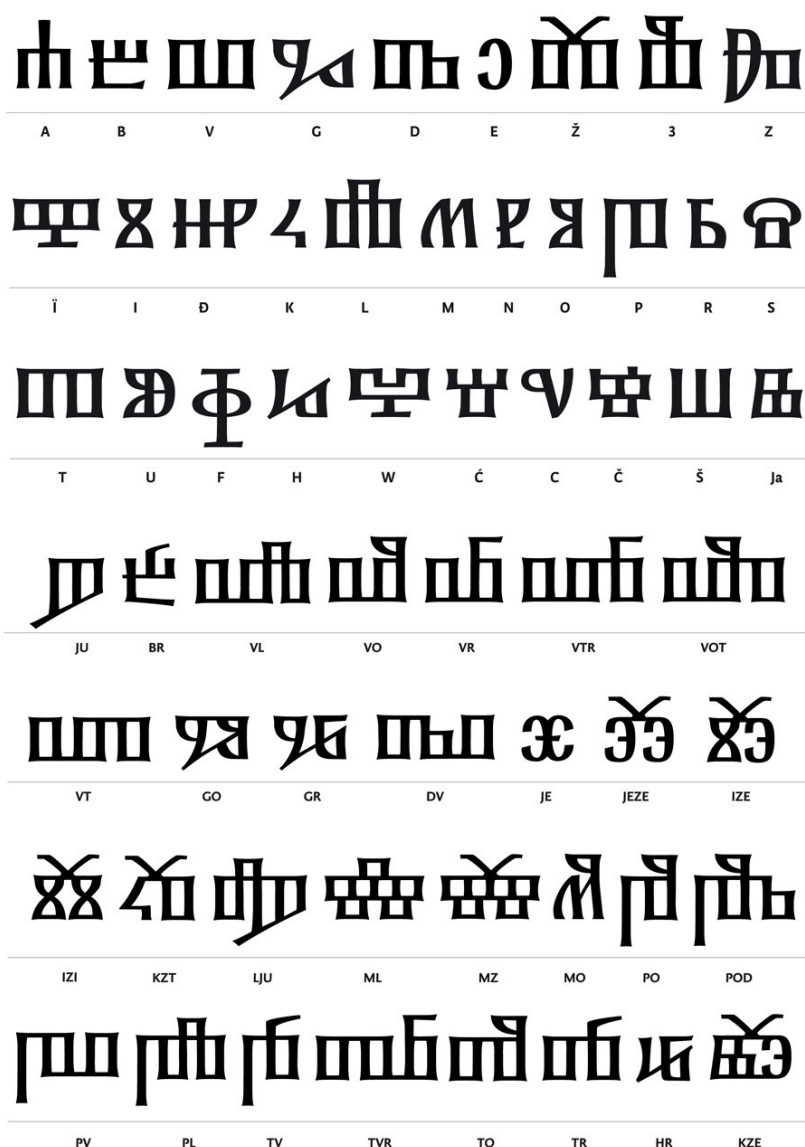


FIGURE 3. Angular Glagolitic font and its most common ligatures, font recreated by Croatian typographer Nikola Đurek (<http://inanutshell.hr/en/exhibits/typography/glagolitic.html>, 26.9.2022.)

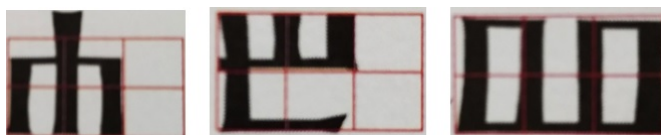


FIGURE 4. Examples of letters *a*, *b*, *v* in the angular Glagolitic script's letter module


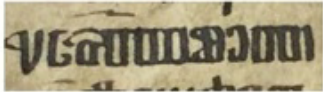



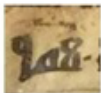



word dividers, etc). Unlike traditional palaeographers, who avoided systematic overviews, the grapholinguistic method takes into account the fact that changes in letter forms take place during changes in the visual transposition of texts. As a result, some patterns between round and angular Glagolitic script have been successfully recognized. Blank space is the main word divider, and its use in 7th-century British and Irish scriptoria marked the beginning of the word separation process (the abandonment of *scriptura continua*). In the context of Glagolitic texts, the use of blank space came after the establishment of angular Glagolitic script (13th century); the more frequent use of initials (for chapter marking), versals (segmentation of text), and punctuation came with it. In angular Glagolitic liturgical texts, punctuation was placed in nearly every blank space in the middle of the line. Another use of punctuations was the marking of numerals, which were (as in Cyrillic script) written as letter forms. Each letter had a numeral value; punctuations combined with a superscripted horizontal line (*titlo*) served as the method by which to differentiate between numerals and standard letters. This practice is common between Glagolitic and Cyrillic scripts, the difference being that Cyrillic script began to use Arabic numerals more often in the 15th century (specifically in diplomatic minuscule script).

The use of abbreviations—suspensions, contractions, and superscriptions (and the aforementioned ligatures)—as a method of saving time and space is another graphetic tool used more often in angular than in rounded Glagolitic script. This data represents yet another interesting fact tested and proven in recent grapholinguistic-based palaeographic research. Although traditional palaeographers recognized this more frequent use of abbreviations, the magnitude of the difference in its usage between the observed scripts became visible after the completion of the aforementioned research.

## 5. Grapholinguistic-Based Palaeographic Research on Croatian Cyrillic Literacy

Unlike Glagolitic script, the origins of Cyrillic script are much easier to explain. Based on Greek uncial script, it developed during the First Bul-

TABLE 1. Examples of abbreviations (contractions) from the London fragment of the Saint Apollonia Breviary (Žagar, Badurina, Paskojević 2021: 226)

| Primjer   | Transliteracija<br>i versta kracenja | Pozicija u tekstu   |
|---|--------------------------------------|---|
|    | <i>c(ēsa)r</i><br>Kontrakcija        | Ia: 7. r., Ib: 1. r.,<br>7. r., IIa: 3. r., IIIa:<br>13. r., IIIb: 7. r.                |
|    | <i>c(ēsa)rstvuetb</i><br>Kontrakcija | Ia: 6. r.   |
|    | <i>b(og)a</i><br>Kontrakcija         | Ia: 10. r. <i>b(ogo)mb</i> ,<br>Ib: 3. r., 5. r., 14.<br>r. <i>b(ož)e</i> , IIb: 7. r., |
|    | <i>b(r̃st)a</i><br>Kontrakcija       | IIIb: 2. r.<br>IIa 5. r., IIIa 12. r.   |
|    | <i>is(u)b(r̃st)b</i><br>Kontrakcija  | IIa: 2. r., 6. r., IIIb:<br>6. r., 14. r.   |
|   | <i>g(ospod)i</i><br>Kontrakcija      | Ib: 14. r., IIa: 1. r.,<br>IIIa: 6. r., IIIb: 14. r.                                    |
|  | <i>g(ospode)vē</i><br>Kontrakcija    | IIIa: 5. r., IIIb:<br>13. r.  |
|  | <i>g(ospodi)nb</i><br>Kontrakcija    | IIIa: 9. r.   |
|  | <i>zap(o)vēdaū</i><br>Kontrakcija    | Ib: 11. r., IIIa: 2/3.<br>r., 13. r., IIIb: 7. r.                                       |

garian Empire during the reign of Tsar Simeon I the Great (late 9th—early 10th century) and under the cultural influence of the Byzantine Empire. It was most likely created at the Preslav Literary School by students of Saint Cyril and Methodius (Curta 2006: 221-222). The change in the linear system and the script's minusculation in the 14th cen-

ture is the most interesting dynamic observed in my grapholinguistic-based research<sup>7</sup> on the script as regards Mediaeval Croatian literacy. This includes developmental processes in the graphemic system and script of Cyrillic charters and documents from Dubrovnik's Mediaeval Slavic chancery and its significance in the broader geopolitical context, especially in diplomatic correspondence with neighboring countries and principalities (Mediaeval Bosnia, Hum, Duklja, Serbia, various autonomous feudal rulers), as well as with more prominent political entities like the Ottoman Empire. Chronologically, the research covers the period from the late 12th to the late 15th century, following the graphic characteristics of the script used in Dubrovnik's Mediaeval Slavic chancery. The research material included roughly 50 representative charters and documents from the Croatian State Archives in Dubrovnik. This institution has one of the largest collections of Cyrillic documents in the Balkans, counting around 10,000 units (exact number unknown). The focus of the research was on *diplomatic minuscule*, a script mainly used in diplomatic, business, and legal communication (hence the name). This script is graphologically quite different from the widespread *Ustav*, a two-line formal uncial version of Cyrillic script that was mainly used for liturgical purposes. The transition from a two- to a four-line system of text organization—the main characteristic of the minusculation process—conditioned the morphology of a significant number of letters in the script. Morphological changes in the script can be observed from the late 12th century. In the *Charter of Ban Kulin* (1189), a trade agreement between the Banate of Bosnia and the Republic of Ragusa, some letter forms display the beginnings of the morphological characteristics of the diplomatic minuscule. The letter *a* begins to show signs of an elongated main vertical line (the elongated vertical line is the main characteristic of the letter form in the diplomatic minuscule script). This same elongation pattern of the “weak” letter parts is also apparent in the letters *z* and *b*, where the lines drop beneath the lower line. These morphological characteristics marked the beginning of the minusculation process. Elongated letter lines slowly began to disintegrate the two-line schema of the *Ustav*, and the transition to the four-line schema resulted in the development of the new script. The transition to the four-line schema brought changes in the letter module; the horizontal line in the square of the two-line letter module specific to *Ustav* became vertical in diplomatic minuscule. One morphological change that may indicate the usage of the new letter module in *Charter of Ban Kulin* (written in *Ustav*) is the separation of the vertical lines in the letter *k*, which is typical of diplomatic minuscule.

---

7. And also doctoral thesis named “Processes of development of Diplomatic Cyrillic Minuscule in the documents of the Medieval Dubrovnik Chancery.”

The morphological traits of diplomatic minuscule are traced in the charter *Tsar Ioan Asen II grants commercial privileges to the commune of Ragusa (Dubrovnik) and its merchants*. This charter is another trade agreement regulating the trade rights of Ragusan merchants. Written in 1230, it is one of a few surviving documents from the 13th century written in this way. The elongation of letter lines that breach the linear system can be seen in the letters *a, z, r, u, b, c, ê*. Like the *Charter of Ban Kulin*, the letter *k* is written with separated letter lines. This time, the right vertical line lost its angle (and is thus reminiscent of the Latin letter *c*), giving it its recognizable minuscule form.

The standardization of diplomatic minuscule took place in the 14th century. All the main morphological features of the newly created script are apparent in various documents from Dubrovnik and neighboring chanceries (Nemanjići dynasty, chancery of King Tvrtko II of Bosnia). Because of the changes in the linear system, the letter *v* lost its recognizable bellies and took a squared form, while the letter *d* was rotated 90 degrees to the left. Some of the changes that can be related to the influence of Latin are connected to the letters *n* and *č*; *n* was written identically to its Latin counterpart, while the letter *č* took a shape similar to the Latin letter *v*. The last morphological change that can be related to the coordination process took place at the turn of the 15th century, when the clockwise rotation of the letter *b* by 90 degrees simplified its writing in the four-line system. This characteristic can be easily recognized in documents from the two most productive scribes of Dubrovnik's Slavic chancery of that time: Rusko Hristoforović and Nikša Zvijezdić. Also, the morphological characteristic of line elongation is apparent in numerous letters in the diplomatic minuscule script (*g, d, z, ž, i, m, r, u, f, b, ê, c, é*). The main rule was to elongate the lines wherever possible, similar to cursive script. Another graphic characteristic that helps to emphasise the relationship with cursive script is the binding of letters. This feature was not widespread, but it served its purpose—the acceleration of the writing process whenever used. One of the scribes who bound letters frequently is the aforementioned Nikša Zvijezdić.

Most Cyrillic written documents display the usage of abbreviations in all their forms. The frequency of abbreviations (especially ligatures) is relatively low compared to Glagolitic script. The most used abbreviations were superscripts (*ô*), contractions, and a combination of superscripts and contractions. Common examples include some of liturgical names and adjectives like *b(ri)sta, s(ve)tibb, m(i)l(o)stv, g(ospo)d(i)nb*. All these examples belong to the so-called *Nomina Sacra* word group, which is related to the Hebrew tradition of avoiding the pronunciation of God's name (jvhv). This pattern was recognized in the early 20th century by German scientist Ludwig Traube (1907). However, some of these words also have non-religious meanings as well, which is understandable considering the nature of the research material. The absence of ligatures in



TABLE 2. Alphabet of the Diplomatic minuscule script by scribe Nikša Zvijezdić (Žagar, Paskojević 2014: 237)

|   |  |   |          |    |
|---|--|---|----------|----|
| a | z                                      | o | h        | ju |
| b | i                                      | p | ō        | ja |
| v | j izostaje kao samostalni slovni oblik | r | č        | je |
| g | k                                      | s | c        | ks |
| d | l                                      | t | č        | ê  |
| e | m                                      | u | š        |    |
| ž | n                                      | f | <br><br> |    |

the researched Cyrillic corpus may be explained by the fact that, unlike angular Glagolitic script, the morphology of Cyrillic script (specifically diplomatic minuscule) hinders or prevents the creation of a wide variety of ligatures such as those present in angular Glagolitic script. The fact

that minuscule Cyrillic letter forms were simplified and accommodated to faster writing brings additional clarification to this observation.

TABLE 3. Examples of Cyrillic abbreviations, scribe Nikša Zvijezdić (Paskojević 2018: 321)

|   |                                       |                             |
|---|---------------------------------------|-----------------------------|
|    | $\dot{o}^a$                           | Superscript                 |
|    | <i>istinom<sup>u</sup></i>            | Superscript                 |
|    | <i>car<sup>s</sup>ka</i>              | Superscript                 |
|    | <i>svē<sup>s</sup>agō</i>             | Superscript                 |
|    | <i>sasta<sup>le</sup>no</i>           | Superscript                 |
|    | <i>b(o)gu</i>                         | Contraction                 |
|   | <i>g(ospo)<sup>d</sup>(i)nami</i>     | Contraction and superscript |
|  | <i>Tvr<sup>b</sup>'ko</i>             | Ligature and superscript    |
|  | <i>dmitr<sup>o</sup>v<sup>o</sup></i> | Ligature                    |

The use of punctuations, initials, and letters is somewhat similar to Glagolitic script. The most significant difference is that Cyrillic script had a shorter transition to the minusculization process, which resulted in the earlier abandonment of *scriptura continua* as compared to Glagolitic script. Since the Cyrillic research material included only secular documents, initials appear on a much smaller scale as compared to Glagolitic script. Besides numerals, punctuations are also used in the text to end sentences and more extensive text chapters. These are sometimes even

accompanied by the writing of a capital letter, although the orthography was not fully systematized.

## 6. Conclusion

The grapholinguistic-based palaeographic method in the specific (more traditional) context is primarily concerned with analytical palaeography, which mostly focuses on letter morphology (Žagar 2007:54). The defined palaeographic categories, which are the main methodological research tool, broaden this context and emphasise all aspects of the writing process relevant to the efficient transmission of a message from writer to reader—from letter coordination in the linear system to the usage of blank space, abbreviations, and numerals. The large amount of data obtained through grapholinguistic-based palaeographic research can be used in various scientific disciplines (besides history) such as linguistics, typography, orthography, etc. The aforementioned examples are merely representative of the research conducted so far. The intent of the author of the current article was to provide a solid outline of the topic and to promote the use of the grapholinguistic method, which is relatively new in the framework of Croatian Mediaeval palaeography.

## References

- Curta, Florin (2006). *Southeastern Europe in the Middle Ages, 500–1250*. Cambridge: Cambridge University Press.
- Drucker, Johanna (2013). *Intro to Digital Humanities. Introduction*. Los Angeles: UCLA Center for Digital Humanities.
- Eckhardt, Thorvi (1955). “Ustav. Glossen zur paläographischen Terminologie.” In: *Wiener slavistisches Jahrbuch* 4, pp. 130–136.
- Hercigonja, Eduard (2006). *Tropismena i trojezična kultura brvatskoga srednjovjekovlja*. Zagreb: Matica hrvatska.
- Neef, Martin (2015). “Writing systems as modular objects. Proposals for theory design in grapholinguistics.” In: *Open Linguistics* 1, pp. 708–721.
- Novak, Viktor (1952). *Latinska paleografija*. Beograd: Naučna knjiga.
- Paskojević, Kristian (2018). “Razvojni procesi diplomatske ćirilice minuskule u dokumentima srednjovjekovne dubrovačke kancelarije.” PhD thesis. Faculty of Humanities and Social Sciences.
- Schenker, Alexander M. (1995). *The Dawn of Slavic. An Introduction to Slavic Philology*. New Haven: Yale University Press.
- Traube, Ludwig (1907). *Nomina sacra, Versuch einer Geschichte der christlichen Kürzung*. München: Beck.

- Yonchev, Vasil [Йончев, Васил] and Olga Yoncheva [Олга Йончева] (1982). *Древен и съвременен български шрифт [Ancient and Modern Bulgarian Script]*. София [Sofia]: Български художник [B'lgarski hudožnik].
- Žagar, Mateo (2007). *Grafolingvistika srednjovjekovnih tekstova*. Zagreb: Matica hrvatska.
- (2020). "Orthographic Solutions at the Onset of Early Modern Croatian. "An Application of the Grapholinguistic Method"." In: *Advances in Historical Orthography, c. 1500–1800*. Ed. by Marco Condorelli. Cambridge: Cambridge University Press, pp. 176–190.
- (2021). *Introduction to Glagolitic Palaeography*. Heidelberg: Universitätsverlag Winter.
- Žagar, Mateo, Vesna Badurina Stipčević, and Kristian Paskojević (2021). "Londonski odlomak glagoljskoga brevijara o svetoj Apoloniji—prilog tekstološkom, paleografskom i jezičnom opisu." In: *Slovo* 71, pp. 191–240.
- Žagar, Mateo and Kristian Paskojević (2014). "Ćiriličke isprave dubrovačke kancelarije XV. stoljeća između minuskule i kurziva." In: *Filologija* 62, pp. 221–247.

# Reinterpreting the Semiotics of Glagolitic


Katharina Tyran

*Abstract.* In my contribution I am addressing current usage of an archaic Slavic writing system, Glagolitic, in Croatia. Created in the 9th century in the course of Slavic Christianization, *glagoljica* gained traction in mediaeval Croatian territories early on, followed by further independent developments. Although the script historically never gained the status of a persistently widely used system for writing and reading in Croatia, and today apart from small academic circles, hardly anyone can actually read and write it, it is celebrated as a very specific visual sign for national culture. Croatian society however uses the Glagolitic script not for ‘representing’ the language or the spoken word respectively itself, but rather for expressing and marking a specific cultural and ethnic sense of belonging, which I will exemplify by a case example from soccer. *Glagoljica*, I argue, has recently undergone a reinterpretation of its semiotic means. Despite a lack of current referential function as a system for writing and reading, Glagolitic has been conventionalized as an autochthonous national heritage, as a specific sign of Croatian cultural, and thus also national identity. I therefore propose that Glagolitic as a writing system in toto may be grasped as a ‘sign’ and not so much as a system or set of constituent signs, e.g., graphemes, and that it became as such part of a Croatian ‘national knowledge’.

## 1. Introduction

In autumn 2022, on Sunday, September 25, the Croatian national soccer team played against the Austria national soccer team in Vienna’s Ernst-Happel-Stadion. It was a UEFA Nations League soccer match, Croatia won 3-1, and reached the finals, while Austria was relegated. But why open a paper on Grapholinguistics with soccer? How could this sport be probably related to script, writing, graphemes? I initially watched the game at random, but eventually wanted to see it through the end. And while watching the match, I was not at all interested in the game tactics, the Austrians desperately trying to score a second goal, or the Croatian

---

Katharina Tyran  0000-0002-2730-8318  
University of Helsinki, Department of Languages, P.O 24 (Unioninkatu 40), FI-00014  
University of Helsinki  
E-mail: [katharina.tyran@helsinki.fi](mailto:katharina.tyran@helsinki.fi)

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 253–264. <https://doi.org/10.36824/2022-graf-tyra>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

players actually doing so. What caught my eye where the dresses of the Croatian players, where I spotted graphemes and number characters reminding me of the Glagolitic script, an archaic writing system that has been used in parts of Croatia throughout history. But as the players ran and tumbled on the pitch, it was difficult to focus the dresses properly and in detail.<sup>1</sup> At first glance, I was therefore afraid of *Déformation professionnelle*. But with a second glance, strongly focusing on the Croatian team's outfits, I saw that I was right. Croatian midfielder Mateo Kovačić with shirt number 8 did not wear the numerical character <8> on his back, but a Glagolitic <Ѣ>, representing the phoneme /i/ and numeric value 20. Defender Dejan Lovren, who scored the third goal for Croatia that evening with a diving header, wore his number <6>, that Slavic trained eyes will recognize as a stylized Glagolitic <Ѧ> for /r/ and 100. The stylized numerical grapheme was turned upside-down for Andrej Kramarić's back number 9. The first letter of surname of assist to the second goal, winger Ivan Perišić, was not a Latin <P>, but a borrowed Glagolitic <Ѣ>, originally with the phonetic value /n/. <B> in team members surnames such as Brozović, Barišić and Budimir, was replaced by a mirrored Glagolitic <Ѣ>, the grapheme for /o/. Several other graphemes on the players backs reminded of Glagolitic letters, and the designed font for the dress was clearly inspired by the angular Glagolitic script generally. Whoever watched the 2022 FIFA World Cup might have seen this specific font on both the white (home shirt) and blue (away shirt) tricot of the Croatian national soccer representation, too.

What can be detected here, could be approached with an idea of the Berlin based art collective Slavs and Tatars, regularly expressed in one of their lecture performances, "Translitative tease": the desire for an emancipation of sounds from their script (Slavs and Tatars, n.d.). As much as Slavs and Tatars—focusing on Turkic languages in the former Soviet Union—explore the potential of the conversion of script as a part of identity politics, I claim that we can exactly in the Croatian context find such a script, that from the vantage point of the present has emancipated from sounds, where graphemes not long represent phonemes in the first place: Glagolitic in Croatia. We can rather observe identity politics through using an archaic script, with letters emancipated from their initial representation of sounds (or numerical value), as Glagolitic graphemes are transposed in their phonemic value by drawing on the visible similarity with Latin graphemes and therefore becoming a "translitative tease," teasing expected transliterations. Moreover, Glagolitic developed from a set of graphemes to common knowledge, to also pick up the overarching theme of the Grapholinguistics in the 21st century conferences. In the following elaboration, I will dis-

---

1. Details on the game can be found on the UEFA Nations League Homepage (UEFA, 2022).



FIGURE 1. Andrej Kramarić wearing his dress with number 9, a stylized variation of the Glagolitic grapheme <Ћ> turned upside-down and Mateo Kovačić with shirt number represented by Glagolitic <Н>. (Details from official HNS graphics online published on November 28, 2022 on the Facebook-page of HNS (Hrvatski nogometni savez, 2022b))

cuss the reinterpretation of Glagolitic, and I propose to grasp this specific writing system in the Croatian context as a cultural icon (Tyran, 2024) and as a visual reminiscence of national identity. Clearly, nations are imagined communities referring to invented traditions in claiming a common identity (cf. Anderson, 2006; Hobsbawm, 1984). A similar approach regarding the constructedness of icons and iconic notions was presented by Eco, who questions similarity as the main feature of icons, as described by semiotics. He argues for a stronger contextualization in a cultural and historical framework (cf. Eco, 2002, p. 197-230). In my contribution, I approach Glagolitic exactly as such a constructed icon for Croatian national identity, as this grapheme system today proofs to have a widespread impact on many areas of everyday culture and material culture. Glagolitic is enshrined in the common knowledge of Croatian society as a marker for national consciousness and identity, and omnipresent: from soccer to universities, from schools to newspapers, from salami and wine to cravats and dresses, from tattoos to awards—Glagolitic graphemes can be found in numerous contexts (cf. Meyer, 2015; Nazor, 2004; Oštarić, 2018; Tyran, 2019). My approach to Glagolitic is on the intersection of Grapholinguistics and linguistic approaches to writing systems following Coulmas (Coulmas, 1981), Spitzmüller (Spitzmüller, 2013), and Dürscheid (Dürscheid, 2016), who highlight writing as a visual tool for communication besides language, together with the concept of iconicity of script and writing as proposed

in art history by Mersmann (Mersmann, 2015). Based on Derrida's post-structuralist grammatology and the Iconic Turn, it aims to account for the fact that writing and script point beyond language and can thus no longer be studied merely as a linguistic model of communication, but as an iconic medium of proposing new modes of interaction. Mersmann argues for a stronger integration of the cultural context in order to overcome the idealization of alphabetic writing systems as mere representation of sounds. (Mersmann, 2015, p. 13-20; 95ff.)

Such approaches help to better integrate the visual representations of writing and respective meaning to be conveyed. This for sure is important in the context of Glagolitic in present-day usage. It truly is not the only ancient script we can find in contemporary use, I however state that this case is specific as it functions on a national level and both as indexing boundaries to related languages and neighboring nations and strengthening national identity on the inside. And as present as it is in contemporary use, one might argue, it has never been before.

## 2. The History of Glagolitic

The emergence of Slavic writing culture in the 9th Century generally is strongly tied to the apostles to the Slavs Cyril and Methodius, two brothers native to Thessaloniki, the capital at the time of the Macedonian part of the Byzantine Empire. Methodius, the elder brother, was born 812, Cyril (whose given name was Constantine) in 826 or 827. Both brothers took part in religious and diplomatic missions. Most notably, Cyril and Methodius were chosen to serve as Slavic Christian teachers for missionary work for the Moravian ruler Rastislav in 862. They translated a variety of liturgical texts, prayers and gospels into Old Church Slavonic, the first literary Slavic language which can be classified as a constructed supra-regional Slavic language, based on a South Slavic local idiom (Damjanović, 2002, p. 9-24). At the same time, in 863, Cyril supposedly created the Glagolitic script for these texts' notations (Eckhardt, 1989, p. 32). The original form of *glagoljica* is only reconstructed, first identified written monuments are dated to the 10th century. Such reconstructions assume 36 to 38 hanging and round letters, each also representing a numerical value. In regards of linguistic functionality, the Glagolitic script represents the concept of one grapheme for one phoneme properly. Originally, the script was known under different names, the term Glagolitic or *glagoljica*, as it is designated in Croatian, derives from the verb *glagoljati* (to speak). Similarly, priests using this writing tradition and liturgy in (Old) Church Slavonic tradition and language are called *glagoljaši* (Damjanović, 2002, p. 47-50).

Cyril and Methodius travelled from Thessaloniki to Moravia and later on Pannonia and spread the Old Church Slavonic word and



Glagolitic script among Slavs in these territories. This was a highly political move by that time: It was the explicit wish of Rastislav, the Moravian prince, to christen his pagan subjects in the Slavonic language in order to dilute the strong influence of the German (Salzburg) bishops. Wishing to fulfil his petition, Byzantium conferred the Slavonic apostles and brothers Cyril and Methodius with this task. Subsequently, they 'developed' the Old Church Slavonic language and simultaneously a new own script system—the Glagolitic—within which many scholars find Christian motifs, making the alphabet a *scriptura sacra*, so to speak. This proved a revolution, as it went contrary to the directive of the Trilinguum, which declared that only Latin, Hebrew and Greek could be used as liturgical languages. Being accused of heresy, Cyril and Methodius travelled to Rome, where the pope recognized their efforts and allowed for Old Church Slavonic and Glagolitic to be used in liturgical concerns. Following the deaths of Cyril and Methodius, their pupils and followers however were expelled from Moravia and Pannonia, with at least some returning to the Balkan Peninsula. This led to an expansion of the Old Church Slavonic language and writing culture into the South Slavic area (Damjanović, 2002, p.9-24). Here now, Glagolitic had to concur with Cyrillic, which had developed based on the Greek uncial from the end of the 9th century and was in use as the official script in the Bulgarian empire, with its capital Preslav. Subsequently, many texts that had been written only in Glagolitic script were transcribed into Cyrillic. In other places, however, most notably Ochrid (today North Mazedonia), scholars stuck to the Glagolitic tradition. (cf. Damjanović, 2002, p. 50-52) Yet, in the South Slavic territories of Orthodox faith and under Byzantine leverage, Glagolitic lost ground and was replaced by Cyrillic.

In Croatia, however, that was part of the *Slavia Latina*, the Glagolitic script gained traction early on, followed quickly by further independent developments, such as the transformation of the originally round form of the Glagolitic graphemes to an angular form. One of the most famous Croatian medieval written historical monuments dated to the 11th century, the *Bašćanska ploča* (the Baška tablet), is already carved in a transitional type of the round to angular Glagolitic script. The Baška tablet is a limestone of almost  $2 \times 1$  meters, with an inscription of 13 rows, and a deed of donation as regards content. This written monument in itself became a famous motif, reproduced countless times in several sizes and as several objects such as magnets or posters. Towards the end of the 14th century, we also find the development of a cursive form of Glagolitic. Over the time, however, the territory where the Glagolitic script was used increasingly shrank and was with rare exceptions limited to the Croatian coastal lands, Istria and the Kvarner Bay, here mostly in the field of liturgy and religious writing. The first printed book in Glagolitic was a missal from 1483, and Glagolitic was,

however with increasing rarity and territorial and functional limitation, used until the end of the 19th and beginning of the 20th century by individuals. (Eckhardt, 1989, p. 39-49; Nazor, 2004a)

There are three main theories on the origin of the Glagolitic script, that have been disputed in Slavic philology, also drawing on the symbolic implications of Glagolitic graphemes. Firstly, the exogenous theory, also known as the Taylor-Jagić-theory, arguing that the Glagolitic alphabet is a derivative of the cursive Greek script of the 8th and 9th century, with influences coming also from Coptic, Hazar, Syrian, Armenian and other scripts. Critically, this theory concludes that the Glagolitic script could only be the work of one author coming out of the Greek cultural space and knowing many languages, respectively scripts. The second theory is the exogenous-endogenous theory, which claims that some parts of the Glagolitic graphemes are taken from other writing systems whereas other parts are formed out of different, non-linguistic elements. Finally, the endogenous theory, which was supported by the work of Finnish Slavist Černohwostow, states that the Glagolitic alphabet has no precedents in other scripts. He attributes all graphemes to the Christian Symbols of the cross, circle and triangle—with Cyril creating a new script and not leaning on existing ones. (cf. Damjanović, 2002, p. 52-61; Eckhardt, 1989, p. 31-49) Scholars in Slavic palaeography, such as Thorvi Eckhardt, support this theory emphasizing that Cyril did not create the graphemes arbitrarily but rather with a strong creativity and symbolism of individual letters (Eckhardt, 1967, p. 460). Already the first grapheme representing /a/ for instance shows the shape of a cross.

### 3. Becoming National Heritage

Scholarly arguments on the emergence of Glagolitic have clearly focused on questions of originality or eclecticism, symbolic values and intentions as well as taking the lead in claiming Glagolitic as a historic legacy. The latter is specifically important for the Croatian context, where the Glagolitic script has been included in the thesaurus of national identity markers (cf. the concept of Löfgren, 1989), specifically in the course of nation-building processes as Croatia became an independent state following the Yugoslav wars and the break-up of Yugoslavia. Glagolitic has become one of the core symbols for Croatian national heritage. Although the script historically never gained the status of a persistently widely used system for writing and reading in Croatia, and today apart from small academic circles, hardly anyone can actually read and write it, you can hardly travel to or move through Croatia without spotting it in numerous contexts, as mentioned earlier. In these contemporary contexts, however, the representation of phonemes or the readability are the least important. Glagolitic graphemes have emancipated from

the sound, and the social meaning predominates the linguistic meaning of writing.

Having said this, I approach the Glagolitic script in Croatia as a recently strong visible symbol of national identity, also against the background of indexicality (Gal & Irvine, 2019, p. 18). Such indexicalization is possible due to extensive scholarly and semi-scholarly work on the Glagolitic script, respective documents, and traditions. The Zagreb-based research institution *Staroslavenski institut* (Old Church Slavonic Institute) is important to mention in this context. But also semi-academic associations such as the *Društvo prijatelja glagoljice* (Friends of Glagoljica Association) in Zagreb or the *Mala glagoljska akademija* (Small Glagolitic academy) in Roč on the Istrian peninsula are fostering the narrative of Croatian legacy to Glagolitic. Both were established in 1993; the first organizes classes and lectures in schools, libraries and museums on the Glagolitic script; the latter is regularly visited by pupils from all over the country to get the chance to become familiar with the Glagolitic script. School classes attend together with their teachers this academy in summer to learn the history of the script and the literary tradition to which it is tied. They also craft brooches, skirts, shirts and dresses with Glagolitic motifs.

With both a strong academic attention and initiatives in the civic sphere and school Glagolitic is construed as an index distinguishing the Croatian language from its surround, in this very case other South Slavic languages emerging after the split of former Serbo-Croatian as a common language concept. This very specific linguistic situation is the matrix for such indexicalisation, where Glagolitic refers to a certain ideology. Serbo-Croatian has been introduced as a common linguistic concept bridging ethnic affiliations and drawing on South Slavic unity in the second half of the 19th century. After phases of convergence and divergence throughout the 20th century, it eventually broke apart together with Yugoslavia from the 1990s. Since then, four standard varieties have developed out of it, Bosnian, Croatian, Montenegrin and Serbian, and linguists in all four respective countries intensely work on differentiation and also foster discourses on idiosyncratic language history (cf. on this topic for instance Bunčić, 2008; Gröschel, 2009; Neweklowsky, 2010; Okey, 2004; Okuka, 1998). In such discourses, Glagolitic is a possible match that has been strongly emphasized. Drawing on Assmann's concepts of writing culture as cultural memory (Assmann, 2007), there are three important interacting features that can be observed regarding Glagolitic in Croatia: Firstly, the remembering—or orientation towards the past; secondly, it is connected to questions of identity, or political imagination; and third, the cultural continuation, or creation of tradition. Scholarly institutions and academic circles are articulating the Glagolitic script as an ancient cultural legacy, with a 'storyline' dating back to the 9th century and discursively constructed as an ever

since ongoing tradition with constant continuity, and simultaneously presenting themselves as the guardians of such legacy. Such a narrative is used for pronouncing script as a politicized national symbol and marker for identity in a common political imagination (cf. for instance Nazor, 2004b).

Before returning to the soccer dresses and their Glagolitic-inspired font as one of the most recent phenomena regarding the reinterpretation and reuse of the writing system and its graphemes, I would like to highlight another initiative launched from the scholarly community to consolidate Glagolitic as such a national symbol: The introduction of a specific celebration day in support of the Glagolitic script and tradition. This was a quite recent initiative introduced in 2018 by the Institute of Croatian Language and Linguistics, together with other cultural and academic institutions. The petition successfully passed the Croatian parliament in 2019, with the official introduction of “Dan hrvatske glagoljice i glagoljaštva” (Day of the Croatian Glagolitic script and Glagolitic tradition) on 22nd of February. This very specific day was chosen as the Croatian incunabulum and first print in Glagolitic, the missal *Misal po zakonu rimskog dvora* from 1483 was printed on this very day, as the colophon of the missal shows. The missal was discursively positioned in an overarching dispositive of autochthony of Glagolitic in Croatia, and as a unique feature in Croatian history, as it is not only the first print in Glagolitic script and Croatian language, but the first missal in a European context not printed in Latin language and script.<sup>2</sup>

Analyzing the topoi and ideological substrates in statements, explanatory texts and social media posts accompanying the introduction of this specific Celebration day by the included organizations, together with associated illustrations, the construction of a symbolical value and indexicality of Glagolitic can be traced. The central and repeatedly articulated goal of this initiative is to bestow *glagoljica* a specific status in the Croatian society, even if it is no longer used as a script in proper sense—which is even stated clearly. Importantly, political leaders strongly supported the initiative and presented themselves prominently in the media and on social media platforms with products and gimmicks launched for the celebration of *glagoljica* day, such as then Minister of Science and Education Blaženka Divjak, who posed with such an umbrella in her office, and then Croatian President Kolinda Grabar-Kitarović, who used it at official appearances on rainy days. The umbrella shows Glagolitic graphemes jumbled on the surfaces, but no clear written message to be transmitted can be identified.

The *glagoljica* celebration day is however not a stand-alone event, but part of a newly introduced whole month dedicated to the Croatian lan-

---

2. A broader analysis of this initiative can be found in my article on Glagolitic as a cultural icon, that will be published in 2024 (Tyran, 2024).

guage “Mjesec hrvatskoga jezika,” which starts of on February 21st—the International Mother Language Day—followed by the *glagoljica* celebration day and finishing of on March 17th, as on this very day, 1967 Croatian linguists, philologists and academics published a Declaration on the Name and Status of the Croatian Literary Language [*Deklaracija o nazivu i položaju hrvatskog književnog jezika*], declaring in favor of an autonomous language concept and glossonym besides Serbo-Croatian. This is the specific conceptual combination important for building up and positioning the Glagolitic script in Croatian society, as an index of differentiation, as a distinguishing feature from the neighboring legitimization and planning of new national languages out of Serbo-Croatian, which there meanwhile are now four—Bosnian, Croatian, Montenegrin, Serbian—are accelerated. National academic institutions work at high pressure on every boundary to either of the other varieties, which in regard to the Croatian language also includes references to *glagoljica* and respective writing tradition. It is exactly the strong visible recognition value of script that is beneficial, too.

This distinctive visible recognition value was picked up by the designers for the Croatian soccer team’s dress, as outlined in the introduction, and turned into a specific font for this purpose. As Spitzmüller has highlighted, typography is part of grapholinguistics, as the materiality of communication does have an impact on the message conveyed (Spitzmüller in Dürscheid, 2016, p. 209–242). Typography therefore is inherently a semiotic means, and for Glagolitic the semiotic value has been reinterpreted, reshaped and reframed to make it fit in current identity politics in Croatia. When Hrvatski nogometni savez HNS officially presented the new dresses on September 15, 2022, they claimed:

“Croatia is Never Done.

Introducing the 2022 Croatia National Team Collection.

The classic red checks on the Home jersey are remixed with a modern twist to reflect the energy and pride of our country.

The new Away Jersey is inspired by Croatia’s nightlife and natural beauty, with vibrant Laser Blue checks reflecting the vibrancy of our country’s fast-moving festival culture and the azure waters of our coastline.” (Hrvatski nogometni savez, 2022b)

What can be extracted here is a visual significance in legitimizing identity internally as a proud nation full of strength, and to the outside as a vibrant tourist hot spot. The media report of HNS even went further and identified the “passionate, powerful and fiery character of the Croatian nation” represented by the red cheeks on white surface, taking up the coat of arms of the Republic of Croatia, a checkerboard of red and white fields (*šahovnica*), that are interpreted as “globally recognizable symbols of Croatian pride” (Hrvatski nogometni savez, 2022a). The checkerboard pattern is also present on the away dresses, however

not in red and white, but in a lighter and a deeper shade of blue, alluding to the Adriatic Sea. The typeface for numbers and names is indicated as being inspired by the “historic Croatian Glagolitic script” and discursively related to Croatian history and tradition and joins the overarching idea of presenting Croatia in the dynamic of combining legacies of the past and energies of the present (Hrvatski nogometni savez, 2022a).

#### 4. Concluding remarks

The archaic script *glagoljica* represents in a contemporary use a visual representation and icon of linguistic and national identity in Croatia. As such, it does however not have a fixed meaning, but is a variable dependent on context, that is incorporated into prevailing discourse and ideology. Although the writing system has had a rather limited range, geographically as well as functionally, and disappeared as a writing medium for over hundred years now, it is provided with a discourse of tracing back a thousand years of history on Croatian soil. By this it is included in a national master narrative and constructed as one of the specific symbols for autochthony and authenticity in Croatian culture, and therefore also indexes difference. Initiatives such as introducing a celebration day to *glagoljica* and Glagolitic tradition in Croatia, which was thereupon integrated in a whole month dedicated to the Croatian language, strongly foster processes of Glagolitic developing from a graphemic system to what I tend to call “national knowledge”.

As such, Glagolitic apparently lost its linguistic functionality, as the primary task is not transmitting a linguistic message: Graphemes do not necessarily represent phonemes, but a visual idea of “Croatianess”. In the presented case example in soccer for instance, Glagolitic is the underlying pattern for a typographic register, and the microtypographic design level here clearly transmits a message beyond merely the name and number of individual players. The font used on the national soccer representations dresses draws on visual recognition as well as on visual similarity of specific graphemes in Glagolitic and Latin script. It therefore triggers rather association to tradition and a specific historical narrative, of an ‘age-old’ autochthonous Croatian tradition in literacy, writing and culture. It is important in delimitating boundaries in the process of identity formation by linking writing systems to particular ethnic and religious groups. In this way, script becomes a factor as important as language for symbolically expressing and marking cultural identity and affiliation, a denotatum for nation, and ethnicity, and its visualization and materialisation as well.

## References

- Anderson, B. (2006). *Imagined Communities. Reflections on the Origin and Spread of Nationalism*. London, New York: Verso.
- Assmann, J. (2007). *Das kulturelle Gedächtnis. Schrift, Erinnerung und politische Identität in frühen Hochkulturen*. Munich: C.H. Beck.
- Bunčić, D. (2008). "Die (Re-)Nationalisierung der serbokroatischen Standards." In: *Deutsche Beiträge zum 14. Internationalen Slavistenkongress Obriđ 2008*. Munich: Otto Sagner, pp. 89–102.
- Coulmas, F. (1981). *Über Schrift*. 1st ed. Suhrkamp.
- Damjanović, S. (2002). *Slovo iskona. Staroslavenska / starobrvatska čitanka*. Zagreb: Matica hrvatska.
- Dürscheid, C. (2016). *Einführung in die Schriftlinguistik*. 5th ed. Göttingen: Vandenhoeck & Ruprecht.
- Eckhardt, T. (1967). "Die slawischen Alphabete." In: *Studium Generale* 20.8, pp. 457–470.
- (1989). *Azbuka. Versuch einer Einführung in das Studium der slavischen Paläographie*. Wien, Köln: Böhlau.
- Eco, U. (2002). *Einführung in die Semiotik*. Munich: W. Fink.
- Gal, S. and J. T. Irvine (2019). *Signs of Difference. Language and Ideology in Social Life*. Cambridge: Cambridge University Press.
- Gröschel, B. (2009). *Das Serbokroatische zwischen Linguistik und Politik. Mit einer Bibliographie zum postjugoslavischen Sprachenstreit*. Munich: Lincom Europa.
- Hobsbawm, E. (1984). "Introduction. Inventing Tradition." In: *The Invention of Tradition*. Ed. by E. Hobsbawm and T. Ranger. Cambridge: Cambridge University Press, pp. 1–14.
- Hrvatski nogometni savez (2022a). HNS Facebook <https://www.facebook.com/cff.hns/photos>.
- (2022b). "Predstavljani novi dresovi Vatrenih. Svježa energija na prepoznatljivoj hrvatskoj šahovnici." HNS Homepage <https://hns-cff.hr/news/24538/svjeza-energija-na-prepoznatljivoj-hrvatskoj-sahovnici/>.
- Löfgren, O. (1989). "The Nationalization of Culture." In: *Ethnologia Europaea* XIX, pp. 5–24.
- Mersmann, B. (2015). *Schriftikonik. Bildphänomene der Schrift in kultur- und medienkomparativer Perspektive*. Munich: Wilhelm Fink.
- Meyer, A.-M. (2015). "Zum Gebrauch der Glagolica bei der Wiedergabe heutiger Texte (anhand von Tätowierungen und Aufdrucken)." In: *Welt Der Slaven* LX, pp. 166–180.
- Nazor, A. (2004a). "Die glagolitische Schrift." In: *Drei Sprachen, drei Schriften. Kroatische Schriftdenkmäler und Drucke durch Jahrhunderte*. Ed. by S. Lipovčan. Zagreb: Erasmus, pp. 33–49.

- Nazor, A. (2004b). "Die Glagoliza heutzutage." In: *Drei Sprachen, drei Schriften. Kroatische Schriftdenkmäler und Drucke durch Jahrhunderte*. Ed. by S. Lipovčan. Zagreb: Erasmus, pp. 232–237.
- Neweklowsky, G. (2010). *Die südslawischen Standardsprachen*. Vienna: Verl. der Österr. Akad. der Wiss.
- Okey, R. (2004). "Serbian, Croatian, Bosnian? Language and nationality in the lands of former Yugoslavia." In: *East European Quarterly* 38.4, pp. 419–442.
- Okuka, M. (1998). *Eine Sprache—viele Erben. Sprachpolitik als Nationalisierungsinstrument in Ex-Jugoslawien*. Klagenfurt: Wieser Verlag.
- Oštarić, A. (2018). "Commodification of a Forsaken Script. The Glagolitic Script in Contemporary Croatian Material Culture." In: *The Material Culture of Multilingualism*. Ed. by L. Aronin, M. Hornsby, and G. Kiliańska-Przybyło. Vol. 36. Educational Linguistics. Springer, pp. 189–208.
- Slavs and Tatars (2016). "Translitative Tease." <https://slavsandtatars.com/lectures/the-tranny-tease>.
- Spitzmüller, J. (2013). *Graphische Variation als soziale Praxis. Eine soziolinguistische Theorie skripturaler "Sichtbarkeit"*. Berlin: De Gruyter.
- Tyran, K. (2019). "Deutungen kroatischer Schriftkultur. Schreibsysteme als nationale und kulturelle Symbole." In: *Slavic Alphabets and Identities*. Ed. by S. Kempgen and V. S. Tomelleri. Bamberg: Univ. of Bamberg Press, pp. 283–297.
- (2024). "Script as a Cultural Icon. Glagolitic." In: *Iconizing Literature, Art, and Science. Intermediality and Value in Popular Culture*. Ed. by P. Wojcik, H. Höfer, and S. Picard. Palgrave Macmillan.
- UEFA (2022). "Nations League Austria—Croatia." <https://www.uefa.com/uefanationsleague/match/2034552--austria-vs-croatia/>.



# Towards the Integration of Cuneiform in the OntoLex-Lemon Framework



Timo Homburg & Thierry Declerck

*Abstract.* This publication shows our approach to adding representations of graphemes of the cuneiform script into the Ontolex-Lemon model. We define a new vocabulary that adds representations of graphemes and their variants, including etymology and their representations in character description languages. We describe how the ontology model can be generalized to describe graphemes of languages that do not rely on a written script for communication. We then interlink these representations to the Ontolex-Lemon model on one end and, for some instances, to the CIDOC-CRMtex model on the other hand and provide application examples in different scripts.

## 1. Introduction

The Ontolex-Lemon model (McCrae et al., 2017) is used by many big data repositories such as Wikidata (Vrandečić and Krötzsch, 2014) or Babelnet (Navigli and Ponzetto, 2012)<sup>1</sup> to represent lexical information about words, word forms, and their relation to semantic descriptions. Words are often depicted in some kind of writing system, the representations of which may give a researcher additional information about writing styles, different sign variants used to express certain characters and words, and their occurrences. This publication proposes a complementary ontology to the Ontolex-Lemon model, which can capture shapes of cuneiform characters in a semantic web vocabulary. This extension is to be thought of as an extension to represent signs and sign variants of the cuneiform script. Still, it should be understood as so general that it could be applied to other similar typed languages. We

---

Timo Homburg  0000-0002-9499-5840  
i3mainz – Institute for Spatial Information & Surveying Technology, Mainz University  
of Applied Sciences, 55128 Mainz, Germany. E-mail: [timo.homburg@hs-mainz.de](mailto:timo.homburg@hs-mainz.de)  
Thierry Declerck  0000-0002-9450-6648  
DFKI GmbH, Multilinguality, and Language Technology Lab, Saarland University  
Campus D3 2, Germany. E-mail: [declerck@dfki.de](mailto:declerck@dfki.de)

1. See also <https://babelnet.org/> for more details.

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 265–297. <https://doi.org/10.36824/2022-graf-homb>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

envision a second use case of this ontology model to represent sign languages, as described in (Declerck, 2022), but will, for brevity, mainly exemplify the primary use case of representing languages in the cuneiform script.

## 2. Foundations

Cuneiform signs are comprised of cuneiform wedges, which according to (Homburg, 2021) can be described using the following parameters:

- A wedge direction on the unit circle
- An optional wedge size identifier
- Indicators of their shape (e.g., broken, wedge head type, wedge stroke type)

While the cuneiform script itself is part of the Unicode standard and about 900 cuneiform signs<sup>2</sup> are attested, these cuneiform signs may appear in a variety of glyph shapes, which differ in the amount and positioning of the cuneiform wedges. The reasons for these changes in glyph shapes may be different writing styles of the same cuneiform sign in space and time, different habits of scribes of the cuneiform tablets, or possible other explanations concerning the adjacent signs of the respective cuneiform sign on given tablets. This situation is not uncommon in other scripts. For example, in Chinese, differences in the number of different stroke types per Chinese character exist not only traditional Chinese characters and Simplified Chinese characters but also between Chinese characters used in Japanese (Kanji) and in their usage over time (Galambos, 2021; Liang, 2021).

## 3. Related Work

This section discusses related work on linked data dictionaries, character encodings, and data formats common in cuneiform languages used for building the linked data-based character registry.

### 3.1. Linked Data Dictionaries

Linked data dictionaries (Gracia, Kernerman, and Bosque-Gil, 2017) provide, among other benefits, means of connecting words and word forms in written language to concepts in the semantic web, thus allowing natural language processing approaches to extract knowledge from a given textual context more accurately. Linked data dictionaries exist for many languages in well-known data repositories such as Wikidata or Babelnet. For cuneiform languages, the MTAAC (Baker et al., 2017) or ORACC (Tinney and Robson, 2014) corpora provide a suitable basis

---

2. <https://www.unicode.org/charts/PDF/U12000.pdf>










| the<br>elements |  |  |  |  | Sign<br>EME  |
|-----------------|---|---|---|---|--|
|                 |  |  |  |  |  |
| designation     | a   | b   | c   | d   | abc  |
| parameters      | sum of the designations and the indices   |   |   |   | a3 b5 c1   |
| category        | number of elements  |   |   |   | 9 = 3+5+1  |

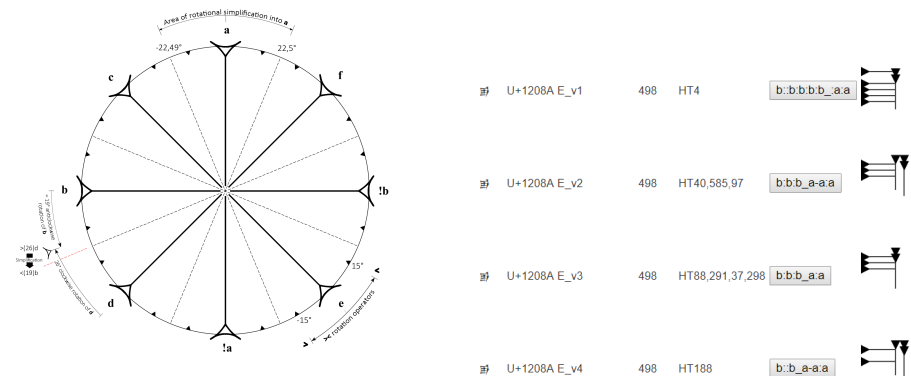
FIGURE 1. Gottstein System for Cuneiform signs from (Gottstein, 2013)

for the extraction of linked data dictionaries. However, such a process has not been attempted to the author’s knowledge. In the future, we can expect linked data dictionaries to be present for each major language.

3.2. Character Description Languages

For many non-alphabetic languages composed out of strokes, such as Japanese or Chinese, encodings for the description of their character composition have been proposed. The Chinese character description language (Bishop and Cook, 2003a) can compose Chinese characters for font generation. Similar character description languages like KanjiVG<sup>3</sup> exist for Japanese. To the author’s knowledge, fonts for cuneiform languages (Mousavi and Lyashenko, 2017; Piška, 1999; 2008) have been based on either SVG drawings or JPG images of cuneiform signs. Hence, unlike the Chinese character description languages, they have not relied on character description languages to describe their respective cuneiform characters. Images will give an accurate representation of the character in question but do not encode semantic information about the context of the character and its composition—something we deem necessary for a proper digital representation of structured scripts. Character descriptions for cuneiform languages have been attempted by (Panayotov, 2015) and (Homburg, 2019). The Gottstein system for describing cuneiform signs counts the number of wedge types in a cuneiform sign, whereas wedge types are distinguished into four different types, as shown in Figure 1. Sometimes, the Gottstein system is slightly adjusted to define the Winkelhaken wedge (w), i.e., the wedge type with only the wedge head as its distinct type, e.g., in (Homburg, Zwick, Mara, and Bruhn, 2022). PaleoCodage (cf. Figure 2a) aims to capture the structure of cuneiform signs, represent dif-

3. <https://kanjivg.tagaini.net>



(a) PaleoCodage encoding system: Wedge types are assigned to wedges on the unit circle. Operators allow for the modification of wedges for the representation of a certain degree on the unit circle (Homburg, 2021).

(b) Sign variants of the same cuneiform sign E in the same space and time and found in the same location and described with different PaleoCodes

FIGURE 2. PaleoCodage encoding system and sign variants of the same cuneiform sign E

ferent sizes of cuneiform wedges, and aims to capture repetitions of substructures of cuneiform signs. This enables PaleoCodage to accurately model cuneiform sign variants even in the same spatiotemporal context as shown in Figure 2b. Given two established character description languages for the cuneiform script, cuneiform characters can be described with two different goals in mind: To index them per cuneiform wedge types (Gottstein) and to describe their shape using PaleoCodage. Both representations may, to a certain extent, be convertible to RDF and, depending on the needs of respective scholars, can serve as a basis for querying different features of these abstracted representations of cuneiform signs.

### 3.3. ATF and JTF

To transliterate cuneiform tablets, two main transliteration formats exist. The ASCII Transliteration Format (ATF)<sup>4</sup> is the primary format of distribution of cuneiform transliterations for all cuneiform languages and exists in many different dialects and varieties which often differ per repository. JTF<sup>5</sup> is a JSON format (Bray, 2017) that includes the same elements as ATF but in a better machine-processable and extendable

4. <http://oracc.museum.upenn.edu/doc/help/editinginatf/cdliatf/index.html>

5. <https://idcs.hypotheses.org/234>

format. It is currently adopted by the Cuneiform Digital Library Initiative (CDLI)<sup>6</sup> and possibly other repositories as a storage format for cuneiform transliterations. Cuneiform transliterations can be rendered from JTF to ATF so that JTF does not provide a replacement format for ATF. Given these two common transliteration formats for cuneiform language transliterations, the JTF format seems to be suited to be extendable for linked data, as defining a JSON-LD context is an easy way to create compatibility with the ontology model we define. Both of the aforementioned transliteration formats do not provide support for paleographic descriptions in any way.

#### 4. Extending the OntoLex-Lemon Model for Cuneiform Paleography

This section outlines our approach for integrating the cuneiform script into the OntoLex-Lemon model. At first, we introduce some terminology we use in our ontology model in Section 4.1, then describe the digital representation of a character in cuneiform languages in Sections 4.2 and 4.3 and how to represent its composition in Section 5.2. After discussing the relation of characters in the ontology model to OntoLex-Lemon Section 4.5, we focus on the description of relations, shape, and provenance of different graphemes by introducing a comprehensive paleographic description vocabulary Section 5. Finally we discuss the integration of etymology concepts in Section 6 and conclude the description of the ontology model by introducing terms to describe glyph occurrences Section 6.2.

##### 4.1. Preliminary Definitions

In order to define a vocabulary for describing characters, we would first like to define certain terms that will be used throughout this publication. These definitions are intended to be so general that they may also be applicable to other languages with similar scripts.

**DEFINITION 4.1.** – *Glyph: The physical manifestation of a grapheme on a written medium.*

This definition covers written glyphs on any medium and is equivalent to the concept [http://cidoc-crm.org/cidoc-crm/TX9\\_Glyph](http://cidoc-crm.org/cidoc-crm/TX9_Glyph) in CIDOC (Doerr, 2005) CRMtex (Murano and Felicetti, 2021). This would be a single cuneiform sign depicted on a written medium (e.g., a clay tablet) for cuneiform. This cuneiform sign might be a non-standard variant. It

---

6. See <https://cdli.ucla.edu/> for more details.

might deviate from this standard variant because the glyph might be broken and have a different number of wedges or wedges not pointing in the expected directions. For non-written languages, such as sign languages, the ontology model provides a class <http://www.purl.org/graphemon#Movement> to represent, e.g., hand gestures.

**DEFINITION 4.2.** – Grapheme: *Digital representation of relevant features of a representation of a glyph or equivalent non-written representation.*

A grapheme represents an idealized or canonical form of a set of glyphs, represented by a digital representation, i.e., abstraction of the set of glyphs describing the cuneiform sign and may be described by an identifier such as a Unicode code point or a dictionary entry number.

**DEFINITION 4.3.** – GraphemeVariant: *A variant of a Grapheme that is associated with the same Unicode codepoint or a semantically equivalent identifier and other identifiers but differs in its normalized visual appearance.*

A <http://www.purl.org/graphemon#GraphemeVariant> is usually connected to a variety of Glyph instances that represent the respective Grapheme variant on physical artifacts in space and time.

**DEFINITION 4.4.** – GraphemeManifestation: *The manifestation of a grapheme either on a written medium or using non-written means.*

We define a <http://www.purl.org/graphemon#GraphemeManifestation> as a more general concept for a Glyph. We would like to generalize the ontology model not to exclude, e.g., hand gestures of sign languages that may be represented using video media or representations of spatio-temporal descriptions of positions of movements. As a superclass of GraphemeManifestation we define the class <http://www.purl.org/graphemon#SymbolicRepresentation> to represent all representations which created a symbolic value in any language.

**DEFINITION 4.5.** – GraphemePart: *A representation of a grapheme that is found as a part of some other Graphemes in the same script.*

A <http://www.purl.org/graphemon#GraphemePart> definition relates to parts of characters found in other characters, but also to parts of, e.g., hand gestures that are part of another hand gestures to describe a particular concept. A grapheme part may constitute its own character. If so, it will be represented with its own Grapheme representation, i.e., also be an instance of Grapheme in the linked data graph.

**DEFINITION 4.6.** – AtomicPart: *A representation of an atomic part out of which Graphemes are comprised.*

A <http://www.purl.org/graphemon#AtomicPart> may represent its own meaning, and Graphemes that consist of precisely one atomic part may exist. An example of an atomic part in cuneiform languages would be a single vertical cuneiform wedge which describes the number one in a

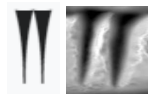
Grapheme. In Chinese, it would be a single stroke that describes a Chinese character (e.g., the horizontal stroke for the number one). However, in Chinese, a horizontal stroke alone might also describe the meaning of horizontal, even though it cannot be used as a Grapheme meaning in this language. In non-written languages, such as in sign language, an atomic part depicts a single unique movement that may be combined with other movements to describe a more sophisticated concept.

## 4.2. What Constitutes a Grapheme?

To describe what constitutes a grapheme in cuneiform languages, we define the following rules, which could also be implemented for automated classification. We assume a cuneiform sign variant to be the standard cuneiform sign variant to represent a particular meaning of a cuneiform sign across time and space. This standard cuneiform sign variant (i.e., its canonical form) might be the most occurring form that the respective linguistic community has agreed upon. It might also be defined per corpus, for example, the most occurring form in a certain corpus. This standard form could be linked to the grapheme data instance, that we define in our knowledge graph. If no such form exists, the grapheme instance in the knowledge graph will simply link to all known grapheme variant instances. For example, consider the cuneiform sign A<sup>7</sup>, which



(a) The cuneiform sign A with its standard form once as grapheme and once as an actual occurrence in the cuneiform text HS 367, front side, column 1, line 3, sign 4



(b) The cuneiform sign A with an alternative form is more common in older cuneiform texts once as grapheme and as an actual representation in HS 1163, back side column 1, line 14, sign 4. This form also resembles the cuneiform sign for the number two 2(disz).

FIGURE 3

constitutes of three vertical cuneiform wedges with at least one attested meaning of water and is described with PaleoCode *a-a:a* shown in Figure 3a. We define a sign variant to A as a variant that differs in one of the following criteria:

7. <https://en.wiktionary.org/wiki/\%F0\%92\%80\%80>

- C.1 Amount of cuneiform wedges per type
- C.2 Positioning of cuneiform wedges towards each other
- C.3 Changes in the type of cuneiform wedges at their respective positions

Figure 3b constitutes such a variant. This example also shows that sign variants may also have the shape of a different standard variant of a sign. In this case, the sign variant of A has the same shape and amount of vertical wedges as the standard variant of sign 2(disz) with the meaning of the number 2.

The definitions also mean that there are differences in cuneiform glyphs that we do not constitute as representing a new sign variant, i.e., a grapheme in the graph structure:

- D.1 The writing order of wedges if known and not exposing a semantic of their own
- D.2 The style of cuneiform wedges themselves (e.g., cuneiform head, cuneiform stroke)
- D.3 The absolute sizes of cuneiform wedges, as long as their proportional size are the same
- D.4 Changes in color or material on which the cuneiform wedges are imprinted unless they capture a semantic meaning

While we deem the latter characteristics not as relevant to distinguish between individual graphemes, they are essential information that should be added to the glyph description in cuneiform languages. Concerning the writing order of wedges, research has started some preliminary work (Taylor, 2014), but has not come to a definite conclusion. However, as long as the writing order of the wedges does not affect criteria C.1-C.3, it is of no relevance for the classification of glyphs as we define in this publication.

### 4.3. Representation of Graphemes in Linked Data

We propose encoding cuneiform graphemes in linked data with two different methods. The first method encodes graphemes using character description language representations like PaleoCodage, the Chinese character description language (Bishop and Cook, 2003b), or the American sign language (Liddell et al., 2003) transliteration as RDF text literals. When no character description languages are available, or as alternative means of definition, SVG literals (Ferraiolo, Jun, and Jackson, 2000) seem to be the natural choice because SVG literals may be displayed in a browser and may serve as the basis for a font generated from the given sign list. Alternative representations might include Open Type Font (Toledo and Rosenberg, 2003) Paths or other image formats such as PNG (Boutell, 1997), which can represent the respective



grapheme. For sign languages, videos or representations of spatiotemporal motions are also viable options. The former may be represented by a hyperlink, the second may use spatial text literals such as Well-Known Text (Herring et al., 2011) in combination with time point extensions. Our ontology model defines literal types for each of these representations.

A second method is to expose the elements that contribute to generating a grapheme directly in RDF. For cuneiform signs, this means that every cuneiform wedge present in a grapheme is represented by its own RDF instance. Hence, a grapheme consists of an RDF subgraph of interconnected AtomicParts. This representation might further semantic exploitation of the individual grapheme but is not practical if queries targeting the grapheme representation only should be answered.

We discuss how to encode a cuneiform sign in RDF using the example of the cuneiform sign A, which we introduced in Section 4.2. The PaleoCode for this grapheme is a-a:a, that is, a vertical wedge *next-to* a vertical wedge *over* a vertical wedge. We can represent this grapheme in RDF as shown in Figure 4. In this example, the grapheme is assigned

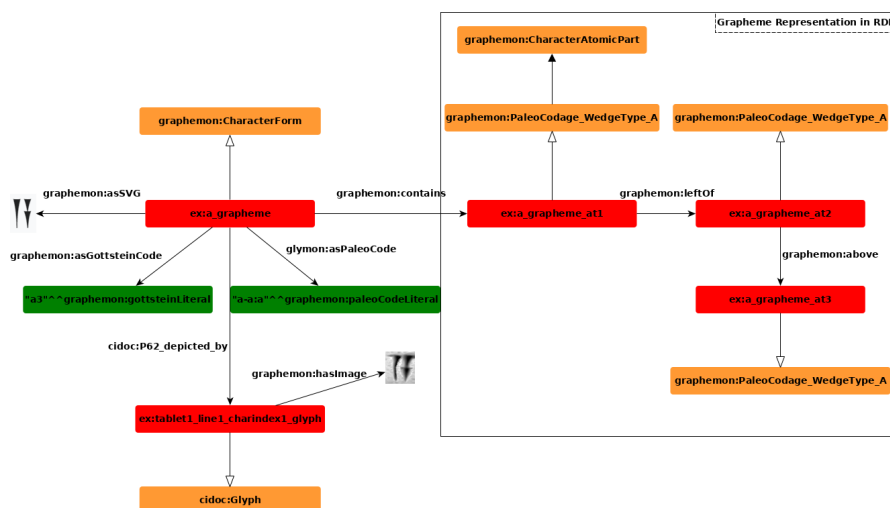


FIGURE 4. Representation of the grapheme structure of cuneiform sign A described with PaleoCode a-a:a

representations in SVG (<http://www.purl.org/graphemon#svgLiteral>), PaleoCodage (<http://www.purl.org/graphemon#paleocodageLiteral>), and in the Gottstein encoding (Panayotov, 2015) (<http://www.purl.org/graphemon#gottsteinLiteral>) and points to a glyph occurrence, while at

the same time, the glyph structure from the PaleoCode is extrapolated in an RDF representation on the right-hand side of the graph. In this RDF representation, even wedge types and atomic parts can be fleshed out in pure RDF. While the literal representations allow querying images of glyphs easily and ready for display by e.g., web browsers, RDF subgraphs may be used to query for clusters of similar representations and also to name such representations in the knowledge graph. Hence, they allow a comparison of shapes of glyphs using the SPARQL query language only, as sets of glyph atomic parts become similarly-shaped subgraphs.

#### 4.4. Grapheme Atomic Parts

In the RDF representation features of single atomic parts, cuneiform wedges could be annotated. That is, each wedge could be annotated with its level of damage or be categorized into a writing style of a different area or scribe. Clearly, this example is only valid for the cuneiform script, and other scripts might include different elements of representation. However, we think these elements could be surmised under a common class structure, which groups similarly styled scripts. For example, Chinese, Japanese, and Cuneiform are all stroke-based scripts, for which the AtomicPart is a stroke of some kind. Figure 5 shows two atomic parts, strokes used in Chinese, the horizontal and the vertical stroke. Both strokes are integral parts of the character for 10, which in itself is included in the word for 11. As an atomic part, the horizontal stroke is also the character for 1, while the vertical stroke is not. Depending on the language, the atomic parts of characters often exhibit a certain order in which they are written. This order may be strict, for example, in Chinese, or it may be superimposed by the encoding used to describe the character variant, such as in the case of cuneiform. To represent a writing order of character atomic parts, these may be described in a <http://www.w3.org/1999/02/22-rdf-syntax-ns#List> or a position vocabulary that we introduce later on in this publication. Sometimes, it may also be sufficient to just state that certain atomic parts are available in a certain grapheme or grapheme part. In this case, a simple <http://www.purl.org/graphemon#partOf> relation is sufficient (cf. Figure 5).

Finally, one might want to capture how the atomic parts of characters are drawn to recreate the abstract character representation for a font. The list of atomic character parts to draw may be appended with positional information extracted from the individual character encoding.

#### 4.5. Connection to Ontolex-Lemon

An important element of this model is its interconnectivity to the Ontolex-Lemon model for modeling semantic dictionaries. To link sign



FIGURE 5. Combination of atomic parts: The *to* strokes, *heng*, and *shu*, which are used to build Chinese characters, are atomic parts and used in the character *shi*, which is used to build the word *shi-yi*, 11.

representations to Ontolex-Lemon word forms, we need to relate components of these word forms to grapheme representations. Unfortunately, an Ontolex-Lemon word form does not have a relation to link to individual graphemes. Instead, it is only possible to link to textual representations of words and word forms as transcriptions, transliterations, or written representations, in essence, represented as text literals. While we cannot change the Ontolex-Lemon model, we can link grapheme instances to instances of word forms described by the Ontolex-Lemon model. To do that, we need to define a new element called `http://www.purl.org/graphemon#WordformOccurrence`, which attests to the representation of a word form with assigned grapheme representations. Figure 6 shows one example connection of Ontolex-Lemon to our ontology model using the word “a,” in its word form “a\_form” and an occurrence of this word form being represented by the grapheme, represented by a variant of the grapheme for the cuneiform sign “A”. In other words, this graph representation allows expressing that a word form can be represented with certain grapheme variant combinations.



TABLE 1. Atomic part description vocabulary for parts of a cuneiform wedge grapheme that represent a semantic meaning and therefore need to be represented in the knowledge graph

| Relation                | Description  |
|-------------------------|--|
| graphemon:angle         | Describes the angle by which the atomic part is rotated if applicable                  |
| graphemon:headColor     | Describes the color of the head of the cuneiform wedge relative to a given scale       |
| graphemon:headSize      | Describes the size of the head of the cuneiform wedge relative to a given scale        |
| graphemon:hasFilledHead | Describes whether the cuneiform wedge head is filled or empty                          |
| graphemon:strokeColor   | Describes the color of the stroke of the cuneiform wedge relative to a given scale     |
| graphemon:strokeSize    | Describes the size of the head of the cuneiform wedge relative to a given scale        |
| graphemon:partStyle     | Describes the style of the cuneiform wedge in a style description language such as CSS |



FIGURE 7. Two different grapheme styles which represent cuneiform signs. The grapheme style with the empty wedge head represents a sign variant present on clay tablets, and the style with the filled wedge head a variant present on stone inscriptions.

on stone rather than on clay (cf. Figure 7). Graphemes of cuneiform signs inscribed on clay usually depict an empty cuneiform wedge head. Therefore, the style in which the grapheme is depicted might in itself contain information about the circumstances in which the grapheme can be found, and useful information to be added to the knowledge graph.

5.2. A Vocabulary for Directions

PaleoCodage and further character description languages relate the different atomic parts of a character to each other by a set of operators and define or reuse an explicit or implicit order of atomic parts. To describe

a cuneiform sign but also further structured scripts in RDF, we formalize these relations in our RDF vocabulary as follows: Individual items may be connected using a set of positional relationships exhibited by the following vocabularies shown in Table 2 to represent the physical relation between atomic parts.

TABLE 2. Relationships between atomic parts: Atomic parts of cuneiform characters

| Relation                | Description  |
|-------------------------|--|
| graphemon:above         | indicates that the current atomic part is above the previous atomic part                 |
| graphemon:below         | indicates that the current atomic part is below the previous atomic part                 |
| graphemon:downright     | indicates that the current atomic part is on the lower right of the previous atomic part |
| graphemon:downleft      | indicates that the current atomic part is on the lower left of the previous atomic part  |
| graphemon:exactPosition | Describes the exact position of the atomic part in a fixed coordinate system             |
| graphemon:left          | indicates that the current atomic part is left of the previous atomic part               |
| graphemon:right         | indicates that the current atomic part is right of the previous atomic part              |
| graphemon:upperright    | indicates that the current atomic part is on the upper right of the previous atomic part |
| graphemon:upperleft     | indicates that the current atomic part is on the upper left of the previous atomic part  |

Table 2 shows the sets of operators we defined to target the cuneiform script. Beginning with a first atomic part, the structure of the cuneiform script follows a subgraph of relations until no such relation can be found. In future work, it may be necessary to define further operators and relations to describe other script types.

### 5.3. Grapheme Relation Vocabulary

Within a script, such as cuneiform, one may encounter parts of individual graphemes reused in other parts of the script. An initial experiment on the representation of all cuneiform Unicode codepoints in one time period-specific font (Homburg, 2021) found that about two-thirds of all cuneiform signs had repeated components in them. Hence, it seems natural to encode these relations in our grapheme description vocabulary so that they can be correlated with, e.g., meanings of the single individual signs and possibly with etymology. When describing the cuneiform

script, we can derive part of individual graphemes from two different sources. The first source may be the definition of the cuneiform signs in standards such as Unicode. For example, the Unicode cuneiform sign AN/AN (AN over AN)<sup>8</sup> is defined by the cuneiform sign AN<sup>9</sup> over another instance of cuneiform sign AN. This makes AN/AN a Grapheme instance which is comprised of two GraphemeParts representing AN. While this definition is used in Unicode, we can generally assume that this definition is not valid for all Graphemes covering all time periods. The reason is that cuneiform signs developed from pictographs and will take the shape on which the Unicode definition is based only at a certain point in time. The second source to derive GraphemeParts from is the representation of Graphemes in a character description language or another structured format. This method was used in (ibid.) and has the distinct advantage that actual representations of GraphemeVariants can be compared by using established and reproducible similarity metric results such as Levenshtein Distance or Image Similarity metrics. In the cuneiform script, as in many other similar scripts, such as Chinese, there are parts of signs that repeat in other signs. This might mean that these signs are related semantically, e.g., that one sign extends a concept introduced by the first sign, that the meaning of two different signs is combined or that the inclusion of one sign in the other has been an artistic choice of the scribe. To model these relations, Table 3 describes properties to express the most occurring types of relations between graphemes. These definitions include two kinds of properties: Properties that derive their conclusions, e.g., from similarity metric calculations, and properties that describe assertions derived by other means. By other means we refer to e.g., the Unicode definition, a scholarly paper or any other external resource which is not readily available and therefore retraceable in the knowledge graph. While these definitions are enough to model relations between cuneiform characters, they might need to be extended for different other scripts.

## 6. Etymology Vocabulary for Graphemes

Etymology is an important concept that helps understand how words have evolved. The Etymological WordNet (De Melo, 2014) showed first how the etymology of words could be traced using a semantic web vocabulary and (Khan, 2018) suggested that the idea of tracing etymology could also be applied to the OntoLex-Lemon model. The resulting ontology model, Etymon (Etymology Model for Ontologies)

---

8. <https://en.wiktionary.org/wiki/\%F0\%92\%80\%AE>

9. <https://en.wiktionary.org/wiki/\%F0\%92\%80\%AD>

TABLE 3. Relation vocabulary between graphemes: Graphemes may be part of other graphemes, modified parts of other graphemes, a generalization, or a combination of other graphemes. Statements like these may stem from metrics or assertions.

| Relation                                | Description   |
|---|---|
| graphemon:isDescribedToBePartOf         | The grapheme is described to be part of the target grapheme             |
| graphemon:isDescribedAsMergedPartOf     | The grapheme is described to be a modified part of the target grapheme  |
| graphemon:isDescribedAsGeneralizationOf | The grapheme is described to be a generalization of the target grapheme |
| graphemon:isDescribedAsModifiedPartOf   | The grapheme is described to be a modified part of the target grapheme  |
| graphemon:isDescribedAsSimplificationOf | The grapheme is described to be a simplification of the target grapheme |
| graphemon:isGeneralizationOf            | The grapheme is a generalized form of the target grapheme               |
| graphemon:isModifiedPartOf              | The grapheme is part of the target grapheme, but slightly altered       |
| graphemon:isMergedPartOf                | The grapheme is merged out of at least two different other graphemes    |
| graphemon:isPartOf                      | Describes the subject grapheme as part of the target grapheme           |
| graphemon:isSimplificationOf            | The grapheme is a simplified form of the target grapheme                |

or lemonETY, describes essential relations and concepts for Etymology that we adjust for the representation of etymology in graphemes. In particular, the concepts <http://lari-datasets.ilc.cnr.it/lemonEty#Cognate>, <http://lari-datasets.ilc.cnr.it/lemonEty#Etymon>, and <http://lari-datasets.ilc.cnr.it/lemonEty#Derivative> are defined in this ontology model. We reuse these concepts in our ontology model but define them on a grapheme level to capture differences in graphemes. Figure 8 shows three examples of an etymological development of cuneiform signs over time. Similar to words, capturing these etymological relations can be of tremendous value for Assyriology research and appropriate machine learning classification tasks. Figure 9 shows how we represent etymology in our ontology model using the example of one cuneiform character in two stages of development. We must stress that the depiction of etymology is just one way to relate grapheme representations to each other. To be precise, etymology describes an inter-
























| Spät-Uruk<br>um 3100  | Djemdet Nasr<br>um 3000   | Frühdyn. III<br>um 2400   | Ur III<br>um 2000   | Altassyrisch<br>um 1900   | Altbabylon.<br>um 1700  | Mittelassy.<br>um 1200  | Neubabylon.<br>um 600   | Archaische<br>Bedeutung     |
|---|---|---|---|---|---|---|---|-----------------------------|
|  |  |  |  |   |  |  |  | SAG „Kopf“                  |
|  |  |  |  |  |  |  |  | NINDA „Ration“              |
|   |  |  |  |   |  |  |  | GU <sub>7</sub> „Zuteilung“ |

FIGURE 8. The etymology of cuneiform characters over time from a pictorial representation to a more abstract representation. Not all representations are depicted by cuneiform wedges. (Labat, 1995)

preted semantic relationship between grapheme representations, even if the semantic is only founded by the sign being a previous or following variant. Another way to represent the similarity between graphemes is to directly exploit their image representations or abstractions thereof.

6.1. Grapheme and Glyph Similarity

Grapheme similarity might be calculated by similarity measures based on either a String representation of the grapheme represented in a sign description language, i.e., a formal textual representation of the glyph depicted, or by a similarity metric based on the pictorial or other representations (e.g., 3D models) of the glyph itself. To enable these kinds of relations in the ontology model, we define one DatatypeProperty score and three base classes <http://www.purl.org/graphemon#SimilarityMetric>, <http://www.purl.org/graphemon#ImageBasedSimilarityMetric>, and <http://www.purl.org/graphemon#StringBasedSimilarityMetric>, from which we might derive script-specific subclasses to express relations between grapheme variants. Table 4.

We recommend using similarity metrics that can be normalized to a percentage range between 0-100 so that comparisons between different similarity metrics can be simplified. However, we do not want to restrict a user from defining arbitrary similarity metric definitions, as long as they are sufficiently documented in the knowledge graph. Given similarity metrics, etymological relationships and assertions about grapheme structures up until the atomic part level, we believe that the relation between graphemes have been sufficiently modeled.

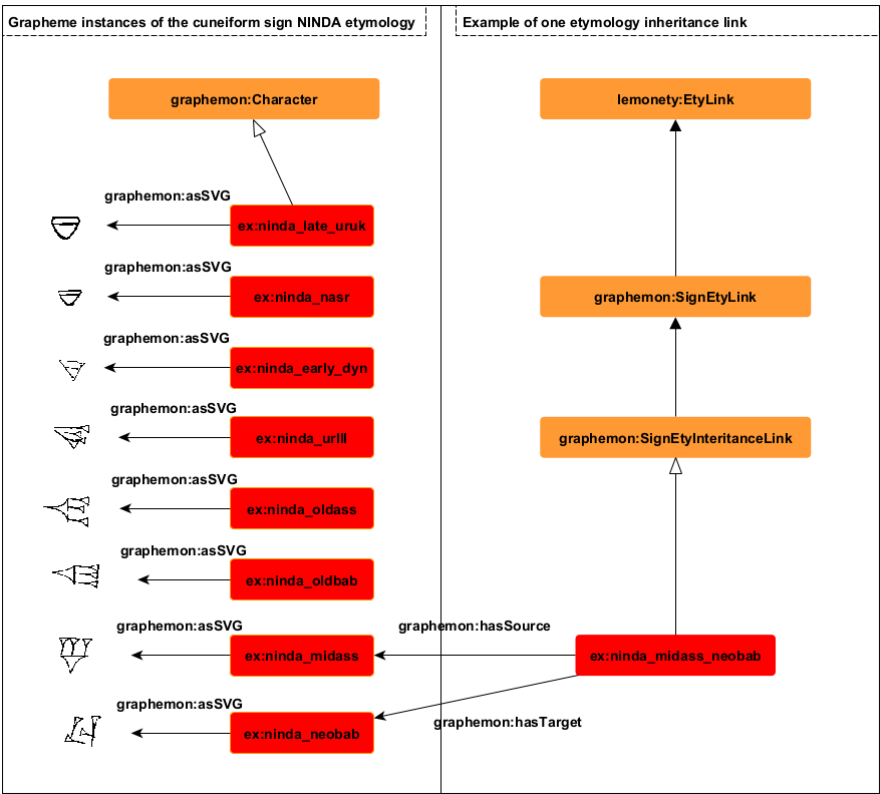


FIGURE 9. Etymology representation of graphemes in the ontology model (only one etymological relation is shown for brevity)

TABLE 4. Classes and properties describing superclasses for similarity metrics and results of similarity metric calculations between grapheme instances or glyph instances

|  |                  |
|--|------------------|
| <code>graphemon:SimilarityMetric</code>            | Class            |
| <code>graphemon:SimilarityMetricResult</code>      | Class            |
| <code>graphemon:ImageBasedSimilarityMetric</code>  | Class            |
| <code>graphemon:StringBasedSimilarityMetric</code> | Class            |
| <code>graphemon:score</code>                       | DatatypeProperty |

## 6.2. Glyph Description Vocabulary

This part of the vocabulary deals with describing visual features of glyph representations. On the example of a cuneiform glyph on a cuneiform

tablet, we will show the aspects of visual representation we deem necessary to be represented in our vocabulary:

- Color representation using the Color Ontology<sup>10</sup> or using CSS literals (<http://www.purl.org/graphemon#cssLiteral>)
- Indicators of damage either on the glyph itself or in its given encoding
- Indicators of the origin of the writers
- Material aspects of the material which was used to represent the glyph
- Metadata of the written script (time period, scribe, etc.)

These vocabulary extensions help identify glyphs by their visual features, another perspective that cuneiform researchers often apply. The Graphemon data model defines the aforementioned properties to be able to model rudimentary features of glyph representations. However, the authors believe that each of these features may be better fleshed out in other vocabularies specializing in the respective fields. Nevertheless, we found it to be a necessity for a researcher to be able to model glyph properties to be able to set them into relation to grapheme representations. In this way, researchers may draw conclusions about the accuracy of the grapheme representations in relation to the given glyph representations.

## 7. Applicability of the Ontology Model for Other Languages

While the ontology model we have proposed is intended for the cuneiform script, we argue that the model also applies to a variety of similarly structured scripts and beyond written languages. We give two examples of written scripts that might benefit from the ontology model and one example of how sign languages, as representatives of non-written languages, can be described using the same or slightly varying terminology.

### 7.1. Egyptian Hieroglyphics and Hieratic Script

Recently, the paleography of Egyptian hieroglyphics and the hieratic written version of these have been digitally captured (Gülden, Krause, and Verhoeven, 2020) and published as a database at the university of Mainz<sup>11</sup>. Databases like these constitute an ideal application case for our ontology model, and this particular database even exposes part of its data as linked open data. As an example of further applicability,

---

10. <https://github.com/timhodson/colourphon-rdf>

11. <https://aku-pal.uni-mainz.de/graphemes>



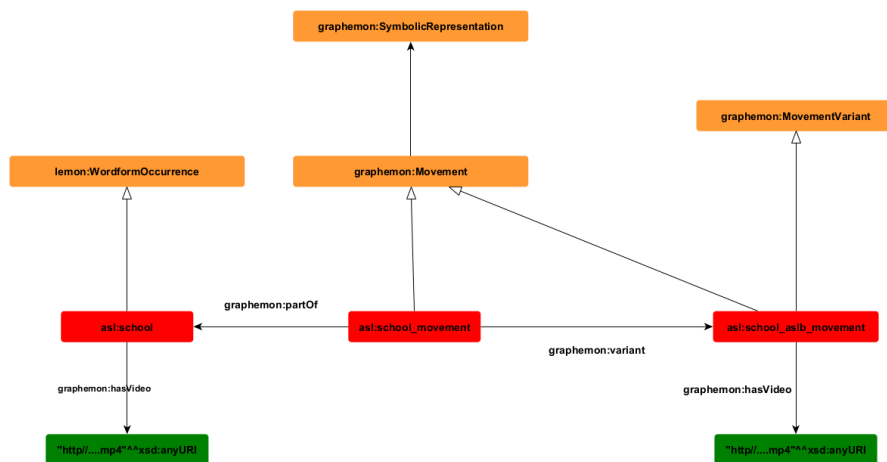


FIGURE 11. Application of the Graphememon model on a hypothetical variant of the American Sign Language (ASL)

of the American Sign language might employ different variants of the base hand gesture, which, in the Graphememon ontology model, would be treated as Grapheme variants.

Figure 11 shows how gestures may be modeled using the Graphememon ontology model. Each gesture becomes an instance of <http://www.purl.org/graphemon#Movement>, an abstract class for gestures. If a sign language like ASL is defined with a standard gesture vocabulary, variants of these gestures become de-facto variants of the initially defined gestures in ASL. As the main topic of this publication is the modeling of written scripts, especially cuneiform, we would like to point out that this part of the ontology model is likely to be fleshed out in future work, as gestures used in sign languages might depict other properties than the written script which will be needed to be modeled as properties in an extended ontology model. Therefore, extensions to the model might likely be developed in future work.

## 8. Application cases

This section discusses the implications of the definition of the ontology model we propose and shows applications in cuneiform studies which directly benefit from its modeling capabilities. In general, we believe that access to structured information about paleography and graphemes, as well as their variants constitutes a missing part in the documentation of primarily digital scholarly editions (Gabler, 2010) of

texts of a different kind. Research on paleography has been done in recent years (Stokes, 2015), and the need for a paleographic vocabulary specific for cuneiform has even been voiced in (Homburg, 2020), but systematic documentation of grapheme variants, their occurrences, and linkage to grammatical forms described by the Ontolex-Lemon model can provide a database to tackle research questions which combine questions of linguistics and paleographic research, an area which is sought to be better understood in a variety of languages (e.g., Maya language, hieratic script, cuneiform, Chinese). In the following, we exemplify immediate application cases enabled by the ontology model with a specific emphasis on cuneiform languages.

### 8.1. A Cuneiform Sign Variant Registry

A cuneiform sign variant registry is a web-based repository that allows the registration of grapheme variants of cuneiform signs, including its spatio-temporal context and further attributes. It attests these variants in different cuneiform transliterations, in different cuneiform languages, and on different cuneiform artifacts. The data structure we propose can be seen as the foundation of such a sign variant registry, which, apart from the functionality of encoding signs, might also help Assyriologists to search for a particular grapheme in its spatio-temporal contexts and find representations of this Grapheme as actual glyph image representations. In essence, the cuneiform sign registry needs to be able to store:

- GraphemeVariants described with unique identifiers and accompanied with metadata:
  - Spatiotemporal context
  - Attested cuneiform language
  - Etymology mappings
  - References to texts or URIs to annotations that describe the sign variant
- Sign definition as an image or in a character description language
- Search indices as similarity metric results between cuneiform signs

Figure 12 shows a screenshot of the JavaScript test tool for PaleoCodage. It can create cuneiform sign variants by entering the character description language code and stores already entered PaleoCodes in a git repository. The repository contents may be downloaded in an RDF representation. An extended version of this PaleoCode storage with support for etymology, cuneiform languages, textual references, and further metadata based on the Graphemon Ontology model is our vision for storage and good accessibility of cuneiform sign variants. The architecture of this repository already fulfills the criteria to represent cuneiform sign

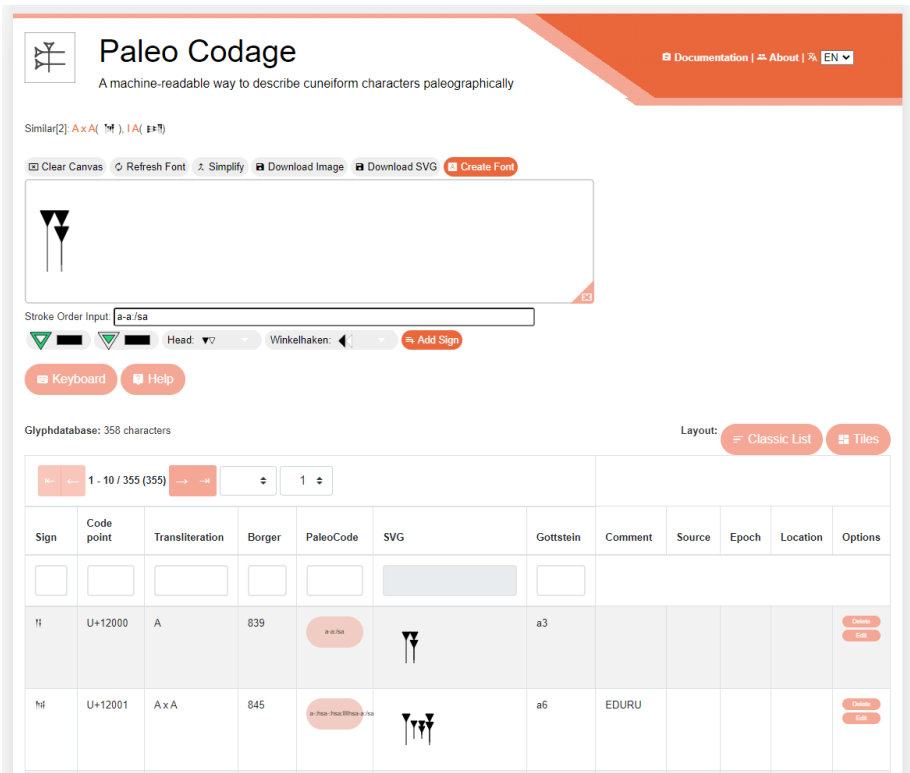


FIGURE 12. A precursor of a cuneiform sign registry which may be extended with the Graphemon ontology model as a backend

variants and can calculate similarity metrics results between its character representations. We believe it may be applicable to other language types as well.

8.2. Integration of Grapheme Information in Cuneiform Digital Editions

Cuneiform digital edition formats should be able to incorporate Graphemon data, as it is represented in this publication. We, therefore, investigate the suitability of data formats for this purpose and highlight what integration in these formats entails. We thereby have the following assumptions:

1. A cuneiform character variant registry as described in Section 8.1 exists so that graphemes may get their own identifiers (possibly also URIs)
2. The data format should aim to be a single file format for easier portability

The ATF format in any shape does not allow to add annotation information. It does not allow encoding information about character variants without defining yet another ATF dialect such as P-ATF (Homburg, 2021).

```

1 @tablet
  @obverse
3 1. 3(u)_v1

```

LISTING 1. Paleographic extension to the ATF format as suggested by (ibid.). This extension requires unique IDs of graphemes to be defined and used in the actual transliteration text.

Listing 1 shows the proposed P-ATF encoding of (ibid.). Each grapheme is assigned a unique ID used directly in the transliteration. The definition of such IDs is currently arbitrary, as the related work on cuneiform sign variants does not show a universally accepted identifier system for cuneiform signs. While such an identifier could be delivered with the URI or be part of a URI that describes a grapheme, the practicality of usage for the average Assyriologist would be to either use some kind of grapheme autocompletion system dependent on a centralized registry of graphemes or not use yet another dialect of ATF, but rather to treat grapheme variants as text annotations.

The situation differs for TEI/XML-based transliteration representations and JSON-based transliterations such as JTF. TEI/XML allows the representation of glyphs<sup>14</sup> so that links to graphemes and glyph representations as URIs could be drawn. The most promising format, in our opinion, would be a JSON-LD-based representation as an extension of the JTF format.

```

1 {"_class": "object", "type": "tablet", "children": [
2   ....
3   {"_class": "chr", "type": "U+1200", "value": "a", "grapheme"
      : "GRAPHEMEID", "glyphrep": "GLYPHREPRESENTATIONLINK"}
4   ]
5   ....
6   }

```

LISTING 2. JTF format extended to link to grapheme representations

14. <https://tei-c.org/release/doc/tei-p5-doc/de/html/ref-glyph.html>



Listing 2 shows a hypothetical cuneiform tablet transliteration representation in JTF<sup>15</sup>. Somewhere in this transliteration, a character transliterated as *a* is attested and referenced to a Unicode code point. In our ontology model, the Unicode code point may identify the standard Grapheme, e.g., by resolving its URI using a SPARQL query. We add the keys “grapheme” and “glyphrep” to identify the grapheme variant via its URI and to identify a representation of the actual glyph on the cuneiform tablet in one of the literal representations we propose. A picture or another medium might represent this glyph. JTF even allows us to define our grapheme variants in the same file if needed and can easily be related to a JSON-LD context (Sporny et al., 2014) for conversion to a linked data representation. In this way, one could build applications that create new grapheme variants in JTF files, which are later synchronized with a cuneiform sign registry in a repository where the transliteration in JTF is supposed to be stored.

### 8.3. Annotation of Grapheme Variants With Annotorious

Annotorious<sup>16</sup> and Recogito<sup>17</sup> are two open-source annotation libraries in JavaScript which allow for annotations in the W3C Web Annotation Data model (Sanderson, Ciccarese, and Van de Sompel, 2013). The creation of annotations seems like the ideal place to use the Graphemon ontology model. Annotations in linked data are comprised of an annotation target, e.g., an area defined on an image resource and an annotation body. The annotation body describes the annotation information which is attested to the annotation target. Figure 13 shows a customized extension of Annotorious, which creates annotations on images of cuneiform tablet surfaces. The annotation objects created with this tool describe an image area with a PaleoCode and a transliteration string. With both information, a set of cuneiform grapheme variant URIs can be retrieved from the knowledge graph, which the user may confirm. The user may be asked to create and describe a new grapheme variant if no URI can be found. Either way, a URI to describe the selected image area is added to the annotation, making image annotations relatable to Graphemes. This way, an Assyriologist may easily document their Grapheme variants and, using the knowledge graph, find further occurrences of the same Grapheme in other texts for comparison.

---

15. <https://github.com/cdli-gh/jtf-lib>

16. <https://github.com/recogito/annotorious>

17. <https://github.com/recogito/recogito-js>

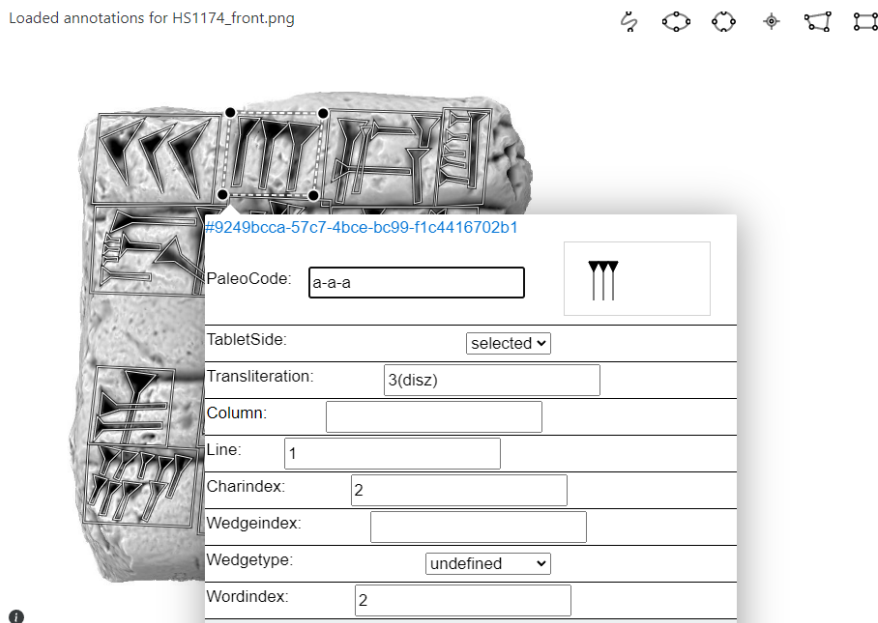


FIGURE 13. Creation of an annotation on a cuneiform 3D rendering using the software Cuneiform Annotator on the MaiCuBeDa dataset (Mara and Homburg, 2023). The marked area denotes the Glyph. The Grapheme is described using a PaleoCode and a Transliteration which can be mapped to a sign name (i.e., a Grapheme representation)

## 8.4. Sample Queries

This section presents sample queries that the new ontology model enables. We show typical applications which are relevant for Assyriology, computational linguistics, and the domain of machine learning.

### 8.4.1. Find Graphemes With Similar Structures

```

1 SELECT ?graphemevariant ?glyphimage WHERE {
2   ?graphemevariant glymon:hasSimilarity ?graphemevariant_sim .
3   ?graphemevariant_sim rdf:type ?PaleoCodeStringSimilarity .
4   ?graphemevariant_sim rdf:value ?simvalue .
5   glymon:hasImage ?glyph .
6   FILTER(?simvalue>0.8)
7 }

```

LISTING 3. A sample query which allows to query cuneiform sign graphemes of similar structure

Listing 3 selects all graphemes above a given similarity threshold of a chosen similarity score. This allows Assyriologists to find similar grapheme variants of cuneiform signs for the sign currently examined and generate similarity statements within the respective text corpus they investigate.

#### 8.4.2. *Etymology of Graphemes*

We can ask for the etymology of graphemes in two different ways and possibly at least two different motivations. The first motivation is to find out about different variants of a grapheme in a specific time period. For example, in Listing 4 we would like to retrieve every Grapheme, including its attested grapheme variants in the Old Babylonian period of cuneiform writing.

```

1 SELECT ?grapheme ?graphemevar ?graphemesvg ?timeperiod WHERE {
   ?grapheme rdf:type cidoc:TX9_Grapheme .
3  ?grapheme graphemon:variant ?graphemevar .
   ?graphemevar graphemon:timeperiod ex:OldBabylonian .
5  ?graphemevar graphemon:as\index{SVG}SVG ?graphemesvg .
   }

```

LISTING 4. Example of querying for etymology relations of a given grapheme

The second motivation is to represent the etymology relations of a given cuneiform sign explicitly and to query similarities between them.

```

1 SELECT ?etymon ?grapheme ?graphemesrc WHERE {
2   ?grapheme rdf:type cidoc:TX9_Grapheme .
   ?grapheme graphemon:variant ?graphemevar .
4   ?etymon graphemon:hasTarget ?grapheme .
   ?etymon graphemon:hasSource ?graphemesrc .
6   }

```

LISTING 5. Example of querying for etymology relations of a given grapheme

Listing 5 queries all graphemes linked in an etymological chain as described in Section 8.4.2. The graphemes can be visualized for assessment by Assyriologists or for extraction by preparation scripts for machine learning analysis.

#### 8.4.3. *Artifacts Including Special Graphemes*

As a third application, we would like to highlight the possibility of different visualizations of grapheme metadata. Similar to already existing approaches such as the cuneiform site index (Rattenborg, 2019), which display cuneiform tablet excavation locations, applications to display the occurrences of specific grapheme variants have not been

present in cuneiform studies. Considering a paleographic enrichment of cuneiform artifact data, one may use the metadata of cuneiform artifacts to create spatial distributions of grapheme occurrences. To achieve this, the Graphemon ontology model needs to be combined with the linked data representations describing the contents of a cuneiform tablet, which can be achieved with the JTF representation presented in Section 8.2. If the knowledge graph includes information on the glyphs on each individual tablet connected to its individual grapheme, each Glyph occurrence can be related to a specific location. Hence, it is possible to create a map representation of glyph occurrences by querying the ontology model.

```

2  SELECT ?grapheme ?graphemevar ?graphemesvg ?timeperiod WHERE {
   ?tablet rdf:type cunei:Tablet .
   ?tablet geo:hasGeometry ?tablet_geom .
4  ?tablet_geom geo:asWKT ?tabgeo .
   ?tablet cunei:contains ?wordformocc .
6  ?wordformocc cunei:contains ?grapheme .
   ?grapheme rdf:type cidoc:TX9_Grapheme .
8  ?grapheme graphemon:variant ?graphemevar .
   ?graphemevar graphemon:as\index{SVG}SVG ?graphemesvg .
10 }
```

LISTING 6. Example of querying for etymology relations of a given grapheme

Listing 6 shows a query returning geocoordinates of findspots of the cuneiform tablets, including a specific GraphemeVariant according to the ontology model. The findspot points can be visualized on a map with an additional indicator as a color, e.g., the time period in which they were found.

A final application case can be discovering and identifying rare graphemes on cuneiform tablets. For this use case, we assume that a sufficiently large corpus of cuneiform tablets has been described using an extended JTF corpus, as described in Section 8.2. One information we can derive from this corpus is the frequency of usage of individual grapheme variants. Rare grapheme variants are graphemes that are not used very often compared to other grapheme variants describing the same grapheme.

```

2  SELECT DISTINCT ?grapheme ?graphemvar COUNT(DISTINCT ?wordformocc AS ?
   graphemvarcount) ?graphemesvg WHERE {
   ?grapheme rdf:type cidoc:TX9_Grapheme .
   ?grapheme graphemon:variant ?graphemevar .
4  ?wordformocc graphemon:contains ?graphemevar .
   }
```

LISTING 7. Example of querying for etymology relations of a given grapheme

Listing 7 states a SPARQL query to retrieve every grapheme with every grapheme variant and an occurrence count of each grapheme variant in the whole corpus. The result may be used as a ranking to retrieve common sign variants and may be combined with other metrics to get an accurate view of their distribution.

## 9. Conclusions

This publication presented a complimentary ontology model to the Ontolex-Lemon model, which can represent graphemes and grapheme variants. This model provides the opportunity to create and contribute to a linked open data cloud of graphemes, glyphs, and signs that can help researchers analyze and discover connections between different visual grapheme representations to classify and retrace the similarities and origins of paleography phenomena. Not only can graphemes be described, but they can also be related to words and actual occurrences of Glyphs, allowing the graph to be used for structured querying, e.g., to obtain instances for targeted machine learning systems. In this way, once enough data has been accumulated, a significant obstacle for machine learning tasks such as sign recognition or cuneiform tablet time period classification (Dencker, Klinkisch, Maul, and Ommer, 2020; Mara and Bogacz, 2019) can be overcome: The acquisition of relevant machine learning data for suitable automation tasks. For Assyriologists, integrating the Graphemon knowledge graph into repositories such as Wikidata, similar to the integration of Ontolex-Lemon for words, would help in documenting, classifying, and including paleographic information in emerging digital editions while at the same time being readily accessible for any data science approaches. We investigated how graphemes may be represented through different media, e.g., character description languages, images, or even videos, in the case of non-written gesture-based languages and how established similarity metrics may compare these different representations. This allows comparing and discovering similar graphemes using different characteristics, which can prove invaluable if sign registries for graphemes are created. Finally, we presented approaches to create, store, manage and query information from the gained knowledge base. The aforementioned components should lead to a better understanding and modeling grapheme variants and cuneiform signs.

### 9.1. Future Work

We see this specification's future work in exploring other scripts and grapheme representations in different languages and consolidating

these results in a working group such as W3C Ontolex<sup>18</sup>. The definition of a unified model for graphemes would allow repositories such as Wikidata to integrate word forms and their semantics in the form of glyph representations. Ideally, we would like to see Wikidata or a similar repository become the data backend of a sign variant registry for cuneiform or any other script that can be modeled in this way. For cuneiform studies, in particular, a formalized knowledge base of this kind is a precious resource for research and retraceability of grapheme variants, and we expect the adoption of these ideas by cuneiform repositories in the future—the adoption of which might pose further research questions and challenges which might need to be addressed.

## Acknowledgements

This paper is based upon work from the COST Action NexusLinguarum—European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology). It is also supported by the Horizon 2020 research and innovation program with the project Prêt-à-LLOD (grant agreement no. 825182).

## References

- Allen, Julie D. et al. (2012). *The Unicode Standard*. Mountain View, CA: The Unicode Consortium.
- Auer, Sören et al. (2007). “Dbpedia: A nucleus for a web of open data.” In: *The Semantic Web*. Springer, pp. 722–735.
- Baker, Heather D et al. (2017). “Machine Translation and Automated Analysis of Cuneiform Languages (MTAAC).” In: *Humanities Commons*.
- Bishop, Tom and Richard Cook (2003a). “A specification for CDL Character Description Language.” In: *Glyph and Typesetting Workshop, Kyoto*. <http://coe21.zinbun.kyoto-u.ac.jp/papers/ws-type-2003/098-cdl.pdf>.
- (2003b). “Character description language CDL: The set of basic CJK unified stroke types.” <https://unicode.org/L2/L2003/03420-cdl-strokes.pdf>.
- Boutell, Thomas (1997). *PNG (Portable Network Graphics) Specification Version 1.0*. RFC 2083.
- Bray, Tim (2017). *The JavaScript Object Notation (JSON) Data Interchange Format*. RFC 8259.

---

18. <https://www.w3.org/community/ontolex/>

- Cidoc, Crm (2003). "The CIDOC Conceptual Reference Model." <http://cidoc.ics.forth.gr>.
- Cyganiak, Richard, Markus Lanthaler, and David Wood (2014). *RDF 1.1 Concepts and Abstract Syntax*. W3C Recommendation. W3C.
- De Melo, Gerard (2014). "Etymological Wordnet: Tracing The History of Words." In: *LREC*, pp. 1148–1154.
- Declerck, Thierry (2022). "Towards a new Ontology for Sign Languages." In: *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 3977–3983.
- Dencker, Tobias et al. (2020). "Deep learning of cuneiform sign detection with weak supervision using transliteration alignment." In: *Plos one* 15.12, e0243039.
- Doerr, Martin (2005). "The CIDOC CRM, an ontological approach to schema heterogeneity." In: *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
- Doerr, Martin, Francesca Murano, and Achille Felicetti (2017). "Definition of the CRMtex." In: [https://www.cidoc-crm.org/crmtex/sites/default/files/CRMtex\\_v1.0\\_March\\_2020.pdf](https://www.cidoc-crm.org/crmtex/sites/default/files/CRMtex_v1.0_March_2020.pdf).
- Eckle-Kohler, Judith, John Philip McCrae, and Christian Chiacros (2015). "LemonUby—A large, interlinked, syntactically-rich lexical resource for ontologies." In: *Semantic Web* 6.4, pp. 371–378.
- Ferraiolo, Jon, Fujisawa Jun, and Dean Jackson (2000). *Scalable vector graphics (SVG) 1.0 specification*. Bloomington: iuniverse.
- Gabler, Hans Walter (2010). "Theorizing the digital scholarly edition." In: *Literature Compass* 7.2, pp. 43–56.
- Galambos, Imre (2021). "Chinese Character Variants in Medieval Dictionaries and Manuscripts." In: ed. by Jörg B. Quenzer, pp. 491–512.
- Gottstein, Norbert (2013). "Ein stringentes Identifikations- und Suchsystem für Keilschriftzeichen." In: *Mitteilungen der Deutschen Orient-Gesellschaft zu Berlin* 145.
- Gracia, Jorge, Ilan Kernerman, and Julia Bosque-Gil (2017). "Toward linked data-native dictionaries." In: *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*, pp. 19–21.
- Greenwade, George D. (1993). "The Comprehensive T<sub>E</sub>X Archive Network (CTAN)." In: *TUGBoat* 14.3, pp. 342–351.
- Gülden, Svenja A, Celia Krause, and Ursula Verhoeven (2020). "Digital palaeography of hieratic." In: *The Oxford Handbook of Egyptian Epigraphy and Paleography*. Oxford: Oxford University Press.
- Herring, John et al. (2011). "Opengis® implementation standard for geographic information-simple feature access-part 1: Common architecture [corrigendum]." In.
- Homburg, Timo (2019). "PaleoCodage—A machine-readable way to describe cuneiform characters paleographically." In: *Proceedings of the Dig-*

- ital Humanities Conference 2019 (DH2019)*, Utrecht, the Netherlands 9-12 July, 2019. Utrecht, Netherlands.
- Homburg, Timo (2020). "Towards Paleographic Linked Open Data (PLOD): A general vocabulary to describe paleographic features." In: *15th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, Ottawa, Canada, July 20-25, 2020, Conference Abstracts*. Ed. by Laura Estill and Jennifer Guiliano.
- (2021). "PaleoCodage—Enhancing machine-readable cuneiform descriptions using a machine-readable paleographic encoding." In: *Digital Scholarship in the Humanities* 36, pp. ii127–ii154.
- Homburg, Timo et al. (2022). "Annotated 3D-Models of Cuneiform Tablets." In: *Journal of Open Archaeology Data* 10.4. ISSN: 2049-1565.
- Kamholz, David, Jonathan Pool, and Susan M Colowick (2014). "PanLex: Building a Resource for Panlingual Lexical Translation." In: *LREC*, pp. 3145–3150.
- Khan, Anas Fahad (2018). "Towards the Representation of Etymological Data on the Semantic Web." In: *Information* 9.12.
- Labat, René (1995). *Manuel d'épigraphie akkadienne. Signes. Syllabaire, Idéogrammes*. Paris: Geuthner.
- Liang, Xiaohong (2021). "An exploratory survey of the graphic variants used in Japan: Part two." In: *Journal of Chinese Writing Systems* 5.2, pp. 115–124.
- Liddell, Scott K. et al. (2003). *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press.
- Mara, Hubert and Bartosz Bogacz (2019). "Breaking the code on broken tablets: The learning challenge for annotated cuneiform script in normalized 2d and 3d datasets." In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, pp. 148–153.
- Mara, Hubert and Timo Homburg (2023). "MaiCuBeDa Hilprecht—Mainz Cuneiform Benchmark Dataset for the Hilprecht Collection." <https://heidata.uni-heidelberg.de/dataset.xhtml?persistentId=doi:10.11588/data/QSNIQ2>.
- McCrae, John P et al. (2017). "The Ontolex-Lemon model: development and applications." In: *Proceedings of eLex 2017 conference*, pp. 19–21.
- Mousavi, Seyed Muhammad Hossein and Vyacheslav Lyashenko (2017). "Extracting old Persian cuneiform font out of noisy images (hand-written or inscription)." In: *2017 10th Iranian Conference on Machine Vision and Image Processing (MVIP)*. IEEE, pp. 241–246.
- Murano, Francesca and Achille Felicetti (2021). "CRMtex-An Ontological Model for Ancient Textual Entities." In: *Decimo convegno annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale, Pisa*.
- Navigli, Roberto and Simone Paolo Ponzetto (2012). "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network." In: *Artificial Intelligence* 193, pp. 217–250.



- Panayotov, Strahil V. (2015). "The Gottstein System Implemented on a Digital Middle and Neo-Assyrian Palaeography." In: *CDLN, London*.
- Piška, Karel (1999). "Fonts for Neo-Assyrian Cuneiform." In: *Proceedings of the EuroTEX'99 Conference*. GUST, pp. 20–24.
- (2008). "Creating cuneiform fonts with MetaType1 and FontForge." In: *TUGboat* 29, pp. 421–425.
- Rattenborg, Rune (2019). "Cuneiform Site Index (CSI): A Gazetteer of Findspots for Cuneiform Texts in the Eastern Mediterranean and the Middle East." <https://ancientworldonline.blogspot.com/2019/12/cuneiform-site-index-csi-gazetteer-of.html>.
- Sanderson, Robert, Paolo Ciccicarese, and Herbert Van de Sompel (2013). "Designing the W3C open annotation data model." In: *Proceedings of the 5th Annual ACM Web Science Conference*, pp. 366–375.
- Seaborne, Andy and Steven Harris (2013). *SPARQL 1.1 Query Language*. W3C Recommendation. W3C.
- Sporny, Manu et al. (2014). "W3C recommendation JSON-LD 1.0."
- Stokes, Peter A. (2015). "Digital approaches to paleography and book history: some challenges, present and future." In: *Frontiers in Digital Humanities* 2, p. 5.
- Taft, Marcus and Kevin Chung (1999). "Using radicals in teaching Chinese characters to second language learners." In: *Psychologia* 42.4, pp. 243–251.
- Taylor, Jon (2014). "Wedge order in cuneiform: A preliminary survey." In: *Proceedings of the 60<sup>e</sup> Rencontre Assyriologique Internationale, Warsaw*, pp. 1–30.
- Tinney, Steve and Eleanor Robson (2014). "Oracc: The open richly annotated cuneiform corpus." <http://oracc.museum.upenn.edu/doc/search/index.html>.
- Toledo, Sivan and Zvika Rosenberg (2003). "Experience with OpenType Font Production." In: *TUGboat* 24.3, pp. 557–568.
- Vrandečić, Denny and Markus Krötzsch (2014). "Wikidata: a free collaborative knowledgebase." In: *Communications of the ACM* 57.10, pp. 78–85.
- Yujian, Li and Liu Bo (2007). "A normalized Levenshtein distance metric." In: *IEEE transactions on pattern analysis and machine intelligence* 29.6, pp. 1091–1095.
- Zesch, Torsten, Christof Müller, and Iryna Gurevych (2008). "Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary." In: *LREC*. Vol. 8. 2008, pp. 1646–1652.



# 16th Century Latin Printed Brevigraphs in Unicode—a Computer Resource

Janusz S. Bień

*Abstract.* A public git repository is presented. It contains some brevigraphs, i.e., specific forms of scribal abbreviations. The brevigraphs are encoded in Unicode. They are organized into two indexes to the scans in the DjVu format: one of the abbreviated word forms and the other one inverted, i.e., of the expanded word forms. From a technical point of view the indexes are just simple CSV files. For browsing indexes, the `djview4poliqarp` program is recommended.

## 1. Introduction


The content of a public git repository<sup>1</sup> is presented, cf. also Fig. 1. The repository contains brevigraphs, i.e., a specific forms of scribal abbreviations (cf., e.g., Honkapohja 2013) found in the two editions of Stanisław Zaborowski's Latin treatise on Polish spelling entitled *Ortographia seu modus recte scribendi et legendi Polonicum idioma quam utilissimus*.

Using git and GitHub for this purpose has several advantages. The interested reader can find easily on the Internet their detailed presentations, I will mention here only the easiness of reporting mistakes and proposing corrections.

The brevigraphs are encoded in Unicode. The standard is not ideal (cf., e.g., Haralambous 2002) but we have to live with it. When needed, private characters are used, proposed by the recommendations of Medieval Unicode Font Initiative added to JunicodeTwo font courtesy of its author Peter S. Baker. The meanings and some other aspects of the brevigraphs are discussed elsewhere, namely in the paper (Janusz S. Bień, 2021).

This resource seems to be the very first computational description of brevigraphs used in printed texts. The primary reference for brevigraphs and other forms of scribal abbreviations is Capelli's *Lexicon abbreviaturarum. Dizionario di abbreviature latine ed italiane* first published in

---

Janusz S. Bień  0000-0001-5006-8183  
retired professor, University of Warsaw, Poland E-mail: jsbien@mimuw.edu.pl

1. <https://github.com/jsbien/Zaborowski-index4djview>

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 299–314. <https://doi.org/10.36824/2022-graf-bien>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

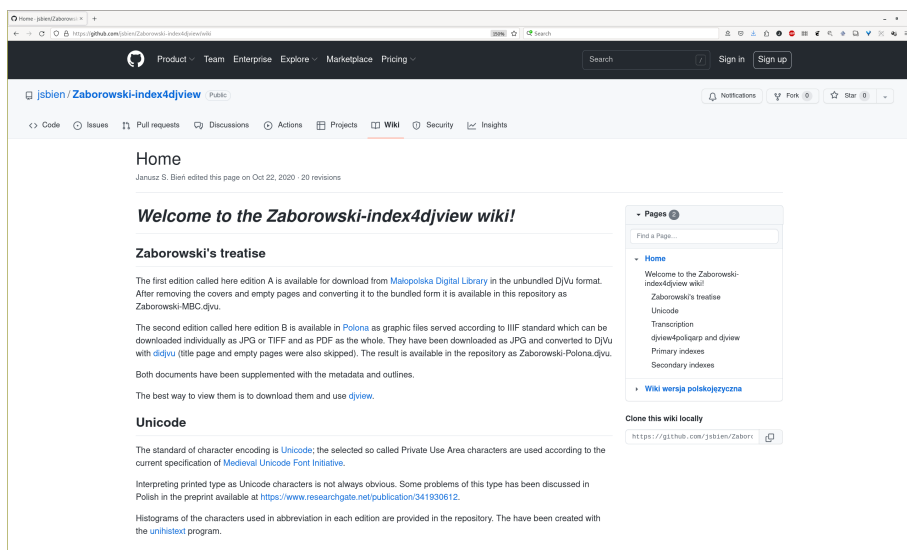


FIGURE 1. Wiki for the repository

1899 (Cappelli, 1889) and available now also as *Capelli Online*<sup>2</sup>. The work however describes handwritten abbreviations, represented by scan snippets, which quite often have no printed equivalents. It's worth noting that the online version was created by crowdsourcing (the call for volunteers was announced in 2015<sup>3</sup>) and the results are freely available also in the source form<sup>4</sup>.

The very first computational approach to Latin abbreviations seems to be Olaf Pluta's Abbreviationes™: A Database of Medieval Latin Abbreviations awarded the 1993 German-Austrian Academic Software Prize (Deutsch-Österreichischer Hochschul-Software-Preis) for outstanding software in the humanities<sup>5</sup>; you can find numerous screenshots in (Pluta, 1995). In 2015 a new version, called Abbreviationes™ Professional, was released. It is said that it

provides a standardized representation of medieval Latin abbreviations by using a Unicode-compliant font (Junicode, created by Peter S. Baker, University of Virginia) which follows the character recommendations of MUFI (Medieval Unicode Font Initiative).

2. <https://www.adfontes.uzh.ch/en/ressourcen/abkuerzungen/cappelli-online>

3. <https://web.archive.org/web/20171015135838/http://www.adfontes.uzh.ch/cappelli/index.php>

4. <https://data.europa.eu/data/datasets/adfontes-cappelli-abbreviaturarum-openglam>

5. <https://olafpluta.net/software/software.html>

To use the software a paid license is required, the price ranges from 99€ for a single fixed IPv4 address to 1199€ for a class B subnet. A trial access is available for free but The author has not used it as he has no intention to purchase this product. The software is mentioned in a chapter of *The Oxford Handbook of Latin Palaeography* (Pluta, 2020).

According to Honkapohja (2021, p. 27) the ORIFLAMMS project made lists of medieval manuscript abbreviations available on GitHub, but I was unable to locate it. Anyway the paper mentions several projects which encode abbreviations and their expansions in the text corpora, but all of them seem to represent manuscripts.

The repository presented here is an open resource. Everybody can use it for any purpose, modify it and distribute modified version etc.

## 2. Printed Texts

The basic notion of the traditional (letterpress) printing is type or sort (both names are confusing because of the ambiguity), i.e., a piece of metal with a face with the (reversed) image of the character to be printed with some appropriate ink on the paper, cf. Fig. 2. A more basic notion is the matrix (a mould) used to cast the types/sorts. For the purpose of encoding we can assume that all types/sorts casted from a single matrix are identical. We propose to call an image of a type/sort on paper, *typoglyph*; even in a single document they can differ because, e.g., of some paper glitch. A generalized, by ignoring such differences, typoglyphs I propose to call typographical characters, in short *typochars* (Janusz S. Bień, 2016–2017 [2019]).

Types/sorts has been kept in the compartments of typescases<sup>6</sup>.

Following (André and Jimenes 2013) the types/sorts put into a single compartment are considered an abstract *typème*. For different type sizes different typescases has been used, so the size is not a property of *typème*, and the same rule holds for some other properties. On the other hand a character with accents is considered a single *typème*, because it is the image of the face of a single type/sort. The notion of typographical character mentioned above has been inspired by the notion of *typème*.

---

6. You can find different cases and their lays/arrangements, e.g., at <http://www.alembicpress.co.uk/Alembicprs/SELCASE.HTM>.



FIGURE 2. A ligature type and its printed image (Daniel Ullrich, Threedots, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=855947>)

### 3. Unicode

#### 3.1. Basics

Let's start this section with a quote from (Korpela, 2006, p. xii): *Character code problems are intrinsically difficult, and very widely misunderstood.*

To make a long story short, Unicode characters have only a loose relation to the characters we use in print or in writing (sometimes called user-perceived characters). One reason is that characters with one more diacritical sign are in principle represented as a base character and the combining characters representing the diacritics; a limited number of characters are available as precomposed ones. In consequence a printed character can be represented by several equivalent Unicode character sequences. Such a sequence is technically called an extended graphemic cluster, which in the author's opinion is a misnomer (a very limited number of these sequences are really graphemic clusters). There is no official Unicode term for the abstraction class of a character object independently of its representation. We propose to call it a *textel* (a text element). However it seems that extended graphemic clusters acquire double meaning: that of a specific sequence and that of a representation independent object. This is how the author understands its use in the SWIFT programming language, cf. also the whole thread on the Unicode mailing list<sup>7</sup>.

#### 3.2. Extending Unicode

In theory Unicode can be easily extended by interested communities by agreeing on the use of private characters. Medieval Unicode Font Initiative mentioned above is a good example. However the private characters are crippled because they are missing the properties provided by

---

7. <https://www.unicode.org/mail-arch/unicode-ml/y2016-m09/0035.html>

the Unicode Character Database<sup>8</sup>. Even if characteristic properties are provided<sup>9</sup>, it may be impossible or prohibitively difficult to make programs to use them; Emacs seem to be an exception<sup>10</sup>.

One of the main Unicode design principles is the distinction between the characters (some abstract objects) and glyphs (their visual images). Another important principle is that Unicode encodes characters, not glyphs. Unfortunately the difference is not always clear.

Let us consider an example. In (Everson et al., 2006, pp. 6, 22, 30, 34, 38) a character was proposed which was used in Latin as an abbreviation for *el*, *ul* and *vel* and in Norse as an abbreviation for *eða*, *el*, *æl* and *al*. The glyphs of this character are presented in Fig. 3. One of them was used as the so called representative glyph in the standard (5.1.0 introduced in 2008) and it served also as the inspiration for the character name.

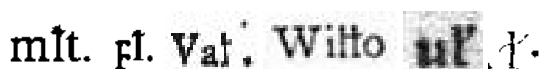


FIGURE 3. Different glyphs of U+A749 LATIN SMALL LETTER L WITH HIGH STROKE

Theoretically there is a method to circumvent these principles by using so called character variation sequences.

The Unicode FAQ<sup>11</sup> contains the following explanation:

Every character in Unicode can be displayed with many different glyphs: An “a” can be displayed with or without the top “hook” (a versus ȁ). A not-equals sign (≠) can be displayed with an angled or vertical slash, and so on.

In some situations, however, it is important to indicate in plain text that only a subset of the possible glyphs for a character should be used, such as a vertical slash for ≠. The variation sequences are a standardized mechanism for requesting such an appearance.

The following FAQ fragment looks even as a recommendation to use variation sequences:

**Q:** I’m proposing an addition to a historic script that is a variant of an existing character. Should I propose it as a new character or as a new variation sequence?

**A:** Variation sequences provide a means to specify a certain significant glyphic variation of a character, without encoding each variation as a separate character. This is particularly useful whenever such distinction is not universally necessary.

8. Cf., e.g., <https://util.unicode.org/UnicodeJsps/character.jsp?a=A749&B1=Show>

9. Cf., e.g., <http://www.kreativekorp.com/charset/PUADATA/>

10. <https://debbugs.gnu.org/cgi/bugreport.cgi?bug=32599>

11. <http://unicode.org/faq/vs.html>

Because the character itself is part of the variation sequence, one should be able to search and find all the instances of that particular character, independent of variation in its appearance, a task which would be more complicated if the variants were encoded as separate characters. If you can replace the variant by the existing character without significantly distorting the content of the text, then a variation sequence is the appropriate way to represent the variant, and you should propose your addition as a variation sequence.

For historic scripts, the variation sequence provides a useful tool, because it can show mistaken or nonce glyphs and relate them to the base character. It can also be used to reflect the views of scholars, who may see the relation between the glyphs and base characters differently. Also, new variation sequences can be added for new variant appearances (and their relation to the base characters) as more evidence is discovered.

The problem is that every usage of variation sequences has to be officially registered by the Unicode Consortium. The only proposal related to Latin scripts the author is aware of, namely (Pentzlin, 2011), has been discussed by the Unicode Technical Committee on a meeting in February 2011 but no action was taken (we are obliged to the author of the proposal for providing this information). The fact does not encourage to submit new proposals.

Several year ago the word *Emojigeddon* was coined, which refers to the flood of emojis accepted into Unicode<sup>12</sup> while the original goals of the standard seem to be neglected<sup>13</sup>. The emojis paved the way for the intensive use of the so called tag characters, e.g., the flag of Scotland is represented as the sequence BLACK FLAG, TAG LATIN SMALL LETTER G, TAG LATIN SMALL LETTER B, TAG LATIN SMALL LETTER S, TAG LATIN SMALL LETTER C, TAG LATIN SMALL LETTER T, CANCEL TAG<sup>14</sup>. Tags have no glyphs, but for the editing and documentation purposes they can be visualised (see below).

In March 2022 Margaret Kibi (marrus-sh) proposed to use tags instead of the variant sequences in the Junicode font<sup>15</sup>. The proposal was supported by several other font users and accepted by the font author. You can find a non-trivial example of this technique in (Janusz S. Bień, 2022b). We hope it will become a kind of a *de facto* standard.

In consequence the important glyph variant of the character discussed above, namely one used in particular in (Balbi, 1460), cf. Fig. 4, which was a book probably typeset by Gutenberg himself, can be en-

---

12. Cf., e.g., <https://www.buzzfeednews.com/article/charliewarzel/inside-emojigeddon-the-fight-over-the-future-of-the-unicode>

13. Cf., e.g., [https://www.explainxkcd.com/wiki/index.php/1953:\\_The\\_History\\_of\\_Unicode](https://www.explainxkcd.com/wiki/index.php/1953:_The_History_of_Unicode)

14. Cf., e.g., <https://emojipedia.org/flag-scotland/>

15. [#discussioncomment-2416880](https://github.com/psb1558/Junicode-font/discussions/122)



FIGURE 4. Giovanni Balbi *Catholicon*, 1460 r

coded as `lSbclfbc` yielding ꝛ (it can be named *l with high stroke ending with flourish*).

Moreover LATIN SMALL LETTER L WITH HIGH STROKE is obviously just a variant of LATIN SMALL LETTER L. Knowledge of this fact can be useful, e.g., for searching and indexing; it is quite surprising that the standard does not provide this information, at least not explicitly. We can now express it formally by encoding ꝛ as `lSbclfbc`. You can find more examples in (Baker, 2022).

The author proposes to call *textons* the instances of Unicode characters which are not used in a self-contained way but are just elements of some sequences. The term was introduced with a slightly different meaning in (Janusz S. Bień, 2016).

### 3.3. Examples of Encoding Problems

Let's have now a look at a specific encoding problem described already in (Janusz S. Bień, 2021).

At first glance the character for *el* seen on Fig. 5 is not available in Unicode.

FIGURE 5. From the left: *vel*, *vel*, *regula*, *populus* (Zaborowski's treatise)

However if you search thoroughly the character proposals stored in the Unicode archive, you will find the proposal mentioned above and learn that this is just a different glyph for U+A749 LATIN SMALL LETTER L WITH HIGH STROKE. It's a pity you cannot find this information directly in the standard.

Let's now have a look at another example, namely the abbreviations of *aliter* and *similiter*, cf. Fig. 6; they come from a recently digitized Zaborowski's treatise edition which has not yet an index, cf. (Janusz S. Bień, 2022a).

The diacritic over *t* looks like a comma. However the character U+0313 COMBINING COMMA ABOVE present in Unicode from its beginning

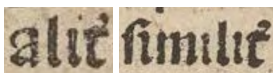


FIGURE 6. From the left: *aliter*, *similiter* (Zaborowski's treatise)

has a different purpose: principally it's the Greek *psili* (smooth breathing mark), although it has some additional applications.

Let us note that Medieval Unicode Font Initiative is not bound by the Unicode rule to encode only characters, not glyphs. So although the Unicode standard has only U+035B COMBINING ZIGZAG ABOVE, in the MUFI specification we can find also U+F1C7 COMBINING ABBREVIATION MARK ZIGZAG ABOVE ANGLE FORM and U+F1C8 COMBINING ABBREVIATION MARK ZIGZAG ABOVE CURLY FORM. So it seems the abbreviations on the Fig. 6 can be encoded as *aliter* and *similiter*; the shapes are not identical to those in the scan, but the distinction is preserved. If you want to avoid the private characters for the reasons described earlier, with the Junicode font you can encode them also as *aliter* and *similiter*.

The last example, cf. Fig. 7, comes again from (Janusz S. Bień, 2021).

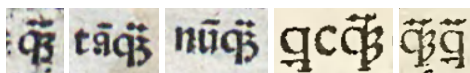


FIGURE 7. From the left: *quam*, *tanquam*, *nunquam*, *quicquam*, *quamquam* (Zaborowski's treatise)

There is no doubt about the characters U+A76B LATIN SMALL LETTER ET added to the standard (together with U+A76A LATIN CAPITAL LETTER ET) in 2008 (version 5.1.0.) following the letter *q*, and the ligature U+E8BF LATIN SMALL LETTER Q LIGATED WITH FINAL ET, a private character present in the MUFI recommendation (Haugen, 2015, p. 79) since version 2. There is also no doubt about U+A757 LATIN SMALL LETTER Q WITH STROKE THROUGH DESCENDER added to the standard (together with U+A756 LATIN CAPITAL LETTER Q WITH STROKE THROUGH DESCENDER) in 2008 (version 5.1.0).

What is problematic here is the encoding of the diacritics. In *tanquam* we have a straight line which can be encoded as U+0305 COMBINING OVERLINE or U+0304 COMBINING MACRON. In *nunquam* the line is not quite straight, is this the same diacritics as in *nunquam* or perhaps a form of tilde (U+0303 COMBINING TILDE)? The words *quam*, *tanquam*, and *nunquam* contain also a diacritic similar to diaeresis, but the dots are more or less connected, is this accidental or intentional? At present the author is not sure what the answer is.

As for *quicquam*, we can assume that LATIN SMALL LETTER ET represents *m* (it used to be written vertically to save the space), hence the meaning of the diacritic is *ua*. In the Unicode archive we can find the document (Everson et al., 2006, s. 8) stating that *ua* can be abbreviated by U+1DD3 COMBINING LATIN SMALL LETTER FLATTENED OPEN A ABOVE. Although the shape of this character in the document is not identical to our example, it seems reasonable to assume this is just a different glyph. This interpretation seem also be confirmed by Erin Blake<sup>16</sup>, who calls the character *jagged horizontal line above letter*. The same diacritic sign occurs twice in an abbreviation, which means *quamquam*; this and other readings quoted here come from (Urbańczyk, 1983, p. 90).

Let us hope that analysing more texts in the future will allow to formulate the definite answers to the questions raised above.

#### 4. DjVu

The format was developed in 1999-2001 for serving scans and underlying text layer over Internet. It's acceptance was hampered by the patents (looks like most of them expired by now), nevertheless it was quite popular in digital libraries, especially in Poland. The open source viewer *djview4*<sup>17</sup> is still actively maintained, cf., e.g., Fig. 8. The DjVu plugin for browser used the now not supported NPAPI interface and no equivalently convenient tool was created. However in my opinion DjVu is still a very good format for scanned documents used offline. One of its advantages is the simplicity of the format.

The DjVuLibre<sup>18</sup> library provides various tools, in particular *djvused* used for operations on the text layer, annotations and metadata.

A typical DjVu document internally contains a dictionary of glyphs (actually connected components), which can be viewed with the *djview4shapes* program<sup>19</sup>, cf. Fig. 9. Another tool for visualisation of connected components dictionaries is Alexander Trufanov's *djvudict* program<sup>20</sup>.

The dictionary of glyphs can be created in particular with *minidjvu-mod*<sup>21</sup> and *minidjvu-mod-gui*<sup>22</sup>.

---

16. <https://collation.folger.edu/2021/09/brevigraphs/>

17. <http://djvu.sourceforge.net/djview4.html>

18. <http://djvu.sourceforge.net/>

19. <https://bitbucket.org/mrudolf/djview-shapes/>

20. <https://github.com/trufanov-nok/djvudict.git>

21. <https://github.com/trufanov-nok/minidjvu-mod>

22. <https://github.com/trufanov-nok/minidjvu-mod-gui>

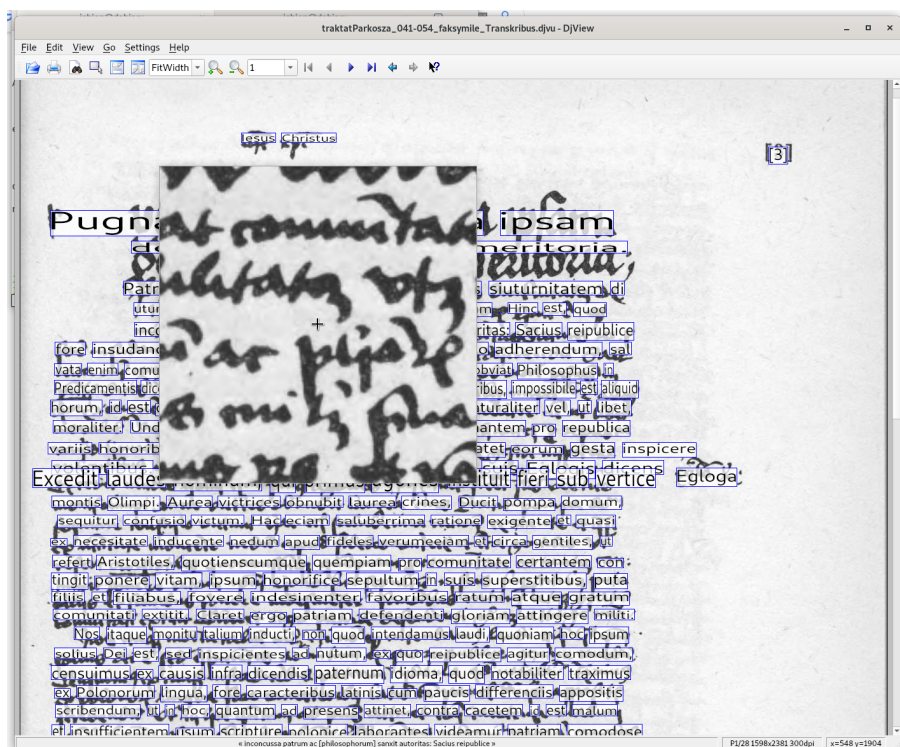


FIGURE 8. djview4: the scan and the underlying text

Looking at the glyph dictionary is the quickest way to get an overview of glyphs used in a text which can help to make the right encoding decisions.

## 5. Indexes to DjVu Documents

The indexed discussed here has been designed by Janusz S. Bień and implemented by Michał Rudolf in the `djview4poliqarp`<sup>23</sup>. From the technical point of view they are just simple CSV files (with the semicolon as the separator). They can be processed in any way a user wishes (we fully agree with Peter Robinson n.d.), but they are most conveniently browsed and edited with the `djview4poliqarp` program mentioned above. The program was originally designed to facilitate creating graphical concordances for the corpora search results, in particular for the

23. <https://bitbucket.org/mrudolf/djview-poliqarp/>

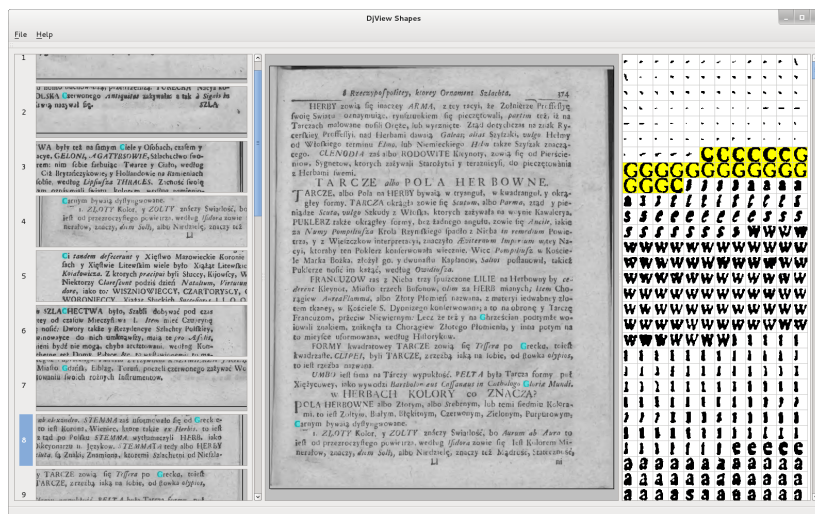


FIGURE 9. djview-shapes: similar shapes grouped together

so called IMPACT Polish Ground Truth corpus<sup>24</sup>, cf., e.g., (Janusz S. Bień, 2014), but was later adapted to handle also offline indexes, cf., e.g., (Janusz S. Bień, 2018a) and (Janusz S. Bień, 2018b), to both online and offline DjVu documents.

The brevigraph indexes discussed here are based on the two editions of Zaborowski's treatise.

The first edition called here edition A is available for download from Małopolska Digital Library<sup>25</sup> in the so called unbundled DjVu format. After removing the covers and empty pages and converting it to the bundled (single file) form it is available in the repository as Zaborowski-MBC.djvu.

The second edition called here edition B is available in Polona digital library<sup>26</sup> as graphic files served according to IIIF standard which can be downloaded individually as JPG or TIFF and as PDF as the whole. They have been downloaded as JPG and converted to DjVu with didjvu<sup>27</sup>; title page and empty pages were also skipped. The result is available in the repository as Zaborowski-Polona.djvu.

Both documents have been supplemented with the metadata describing their origin and the outlines containing traditional page identifiers

24. <https://szukajwslownikach.uw.edu.pl>

25. <http://mbc.malopolska.pl/publication/89609>

26. <https://polona.pl/item/73794330>

27. <http://jwilk.net/software/didjvu>

(the identification of the sheet, the number of the leaf in the sheet, the side *recto* or *verso*).

The best way to view them is to download them and use `djview4` program mentioned earlier.

The indexes are named respectively `ZaborowskiA.csv` and `ZaborowskiB.csv`. We called them the primary indexes.

Every line of an index file consists of three or four fields:

1. The entry used for sorting and incremental search. The entries in the primary indexes consist of the abbreviations, e.g., 'māib<sup>9</sup>'.
2. The reference to the relevant image fragment in the form used by `djview4` viewer mentioned earlier, namely an Universal Resource Locator. In the indexes discussed here the scheme and authority parts are absent, and the path limited to the file names, this means in practise that `djview4poliqarp` has to be called with the index directory as the default one. The fragment part is also missing, and the query part contains the dimensions and the coordinates of the image fragment in the `djview4` specific form; it can contain also the specification of a color used for highlighting. This field is created with an appropriate tool. In particular `djview4` and `djview4poliqarp` can be used for this purpose. Here is an example:

```
Zaborowski_MBC.djvu?djvuopts=&highlight=561,954,133,58&page=1
```

3. A description: a text displayed for the current entry in a small window under the index.
4. An optional comment displayed after the entry. In the primary index this is the abbreviated word, proceeded by REFERENCE MARK for a more distinctive display, e.g., ⌘ *manibus*.

The entries can be displayed in several orders:

- File order, in practise it means the order of the brevigraphs occurrences in the treatise.
- Alphabetic order word by word, i.e., spaces and hyphens are relevant.
- Alphabetic order letter by letter, i.e., spaces and hyphens are ignored.
- *a tergo* (the reverse alphabetical order).

The indexes contain also some additional auxiliary entries.

First of all there are entries describing words which are not abbreviations but are interesting for other reasons; in particular, they document the usage of `LATIN SMALL LETTER ET` as a final 'm'.

Secondly, they are entries allowing, when displaying an index in the file order, to move quickly to a specific page or, when both A and B indexes are displayed together, to the beginning of an edition.

As we have seen already, interpreting printed type as Unicode characters is not always obvious. For verification purposes the histograms

of the Unicode characters used in the abbreviations in each edition are also provided in the repository. They have been created with the unihist-text program<sup>28</sup>.

There are also the secondary indexes named respectively ZaborowskiAi.csv and ZaborowskiBi.csv ('i' meaning 'inverted'). In the secondary indexes the fields 1 and 4 are exchanged, so the abbreviated words are now the entries. They are generated from the primary ones with a sed one-liner program.

Loading both indexes and sorting the joined index in an alphabetic order allows to compare how the words were abbreviated in the A and B editions.

The transcription, based on the one provided by Urbańczyk (1983), has been synchronized with the scans with the use of Transkribus<sup>29</sup>. The results has been used as the basis for indexes, which were later verified and extensively modified. These changes has not been applied yet to the texts stored in the Transkribus system.

In dview4poliqarp program you can use the left panel for displaying the entries you find interesting, cf. Fig. 10 and Fig. 11.

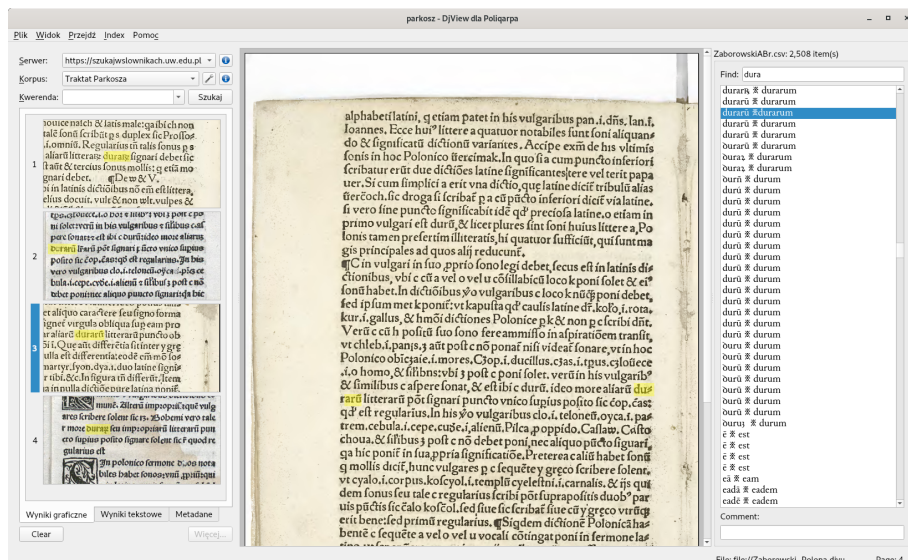


FIGURE 10. A primary index

28. <https://bitbucket.org/jsbjen/unihistext/>

29. <https://readcoop.eu/transkribus/>



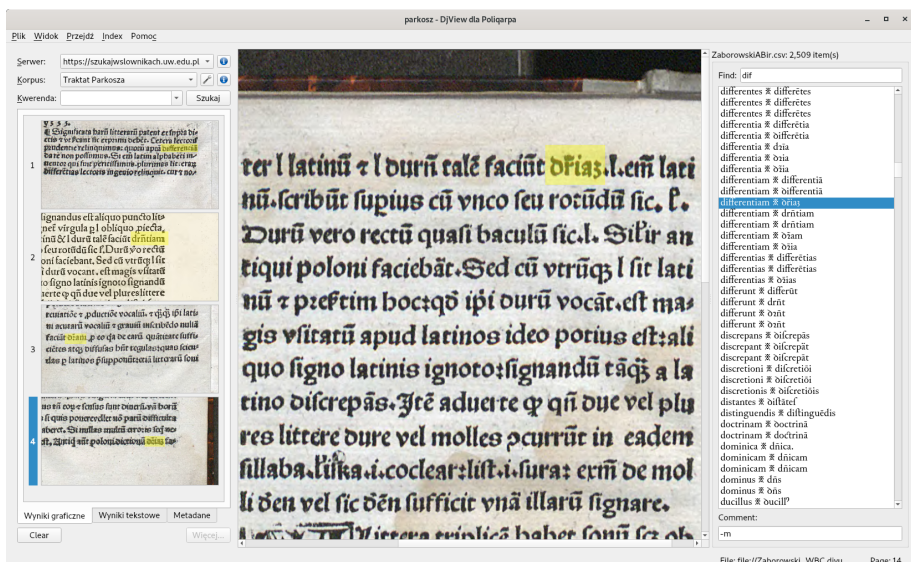


FIGURE 11. A secondary index

As it was mentioned before, the indexes can have other uses beside being browsed. For example, it is quite easy to convert, with just some regular expressions, an index into the djvused input to create a DjVu document where abbreviations are somehow marked and the expansions provided as tooltips, cf. Fig. 12.

## 6. Final Remarks

The paper (Honkapohja, 2021) entitled *Digital Approaches to Manuscript Abbreviations: Where Are We at the Beginning of the 2020s?* was already mentioned earlier. It's main focus is the place of abbreviations in the theory of writing systems, but it contains also a section concerning computer encoding of abbreviations and/or their expansions. With the exception of some early corpora, all the projects mentioned encode the texts in XML, most of them following the recommendations of Text Encoding Initiative<sup>30</sup>, which discusses abbreviations in section 3.6.5<sup>31</sup>.

Perhaps in some future Zaborowski's treatise will be also encoded in TEI XML, but for the present purpose the tools and resources described in the paper are fully adequate.

30. <https://tei-c.org/>

31. <https://tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#CONAAB>



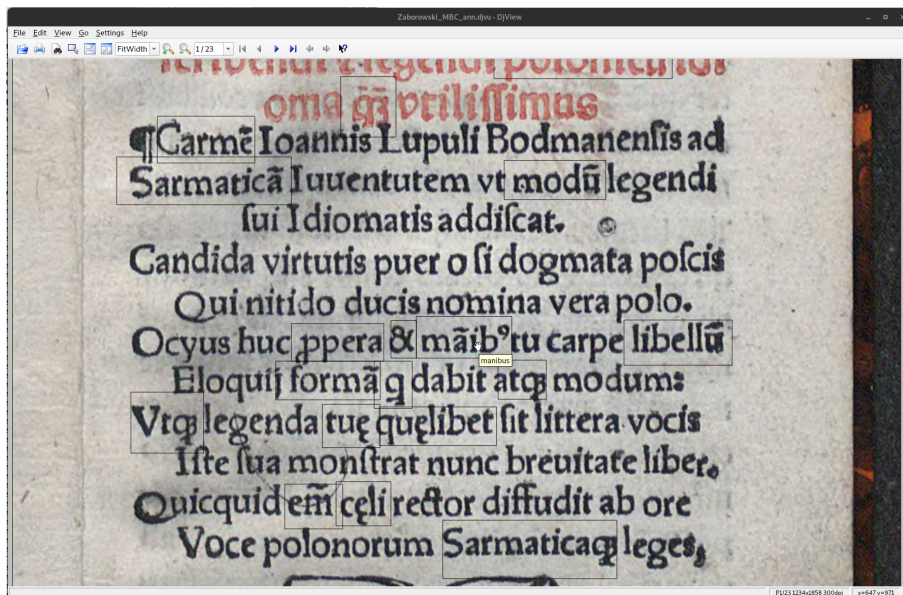


FIGURE 12. A tooltip with the expansion of māib⁹ abbreviation

## References

- André, Jacques and Rémi Jimenes (2013). “Transcription et codage des imprimés de la Renaissance.” In: *Revue des Sciences et Technologies de l’Information—Série Document Numérique* 16.3, pp. 113–139.
- Baker, Peter S. (2022). *Junicode—the font for medievalists. Specimens and user manual for version 2*. <https://github.com/psb1558/Junicode-font>.
- Balbi, Giovanni (1460). *Catholicon*. <https://www.loc.gov/item/47043559/>. Mainz.
- Bień, Janusz S (2018a). “Elektroniczny indeks do słownika Lindego [An electronic index to Linde’s dictionary].” In: *Kwartalnik Językoznawczy* 2015.3, pp. 1–19.
- (2018b). “Elektroniczne indeksy fiszek słownikowych [Electronic indexes for dictionary fiches].” In: *Kwartalnik Językoznawczy* 16.2, pp. 16–27.
- (2022a). “Polskie litery w traktacie Stanisława Zaborowskiego. Litera A i pochodne [Polish letters in Stanisław Zaborowski’s treatise. Letter A and derivatives].” In: *Poznański Półrocznik Językoznawczy* 1, pp. 1–20.
- (2016). “Problemy kodowania znaków w korpusach historycznych [Character encoding problems in historical corpora].” In: *Semantyka a konfrontacja językowa*. Ed. by Danuta Roszko and

- Joanna Satoła-Staśkowiak. Vol. 5. Warszawa: Instytut Slawistyki PAN, pp. 67–76.
- Bień, Janusz S. (2016–2017 [2019]). “Repertuar znaków piśmiennych—problemy i perspektywy [Towards an electronic repertoire of basic text elements].” In: *Kwartalnik Językoznawczy* 2016.2016/4-2017/1, pp. 1–18.
- (2022b). “Representating Parkosz’s alphabet in the Junicode font.” In: *TUGboat* 43.3, pp. 247–251.
- (2014). “The IMPACT project Polish Ground-Truth texts as a DjVu corpus.” In: *Cognitive Studies | Études Cognitives* 14, pp. 75–84.
- (2021). “Traktat Stanisława Zaborowskiego i skróty brachygraficzne [Scribal abbreviations in Zaborowski’s treatise].” In: *Poznański Półrocznik Językoznawczy* 1 (30), pp. 1–42.
- Cappelli, Adriano (1889). *Lexicon abbreviaturarum. Dizionario di abbreviature latine ed italiane*. Milan: Ulrico Hoepli.
- Everson, Michael et al. (2006). *Proposal to add medievalist characters to the UCS*. Tech. rep. N3027. ISO/IEC JTC1/SC2/WG2.
- Haralambous, Yannis (2002). “Unicode et typographie: un amour impossible.” In: *Document numérique* 6.3, pp. 105–137.
- Haugen, Odd Einar, ed. (2015). *MUFI character recommendation version 4.0*. <http://hdl.handle.net/1956/10699>. Medieval Unicode Font Initiative.
- Honkapohja, Alpo (2013). “Manuscript abbreviations in Latin and English: History, typologies and how to tackle them in encoding.” In: *Studies in Variation, Contacts and Change in English. Principles and Practices for the Digital Editing and Annotation of Diachronic Data*. Vol. 14. <https://varieng.helsinki.fi/series/volumes/14/honkapohja/>.
- (July 2021). “Digital Approaches to Manuscript Abbreviations: Where Are We at the Beginning of the 2020s?” In: *Digital Medievalist* 14.
- Korpela, Jukka K. (Jan. 2006). *Unicode Explained*. Sebastopol, CA: O’Reilly.
- Pentzlin, Karl (2011). *Proposal to add Variation Sequences for Latin and Cyrillic letters*. Tech. rep. L2/11-059. ISO/IEC JTC1/SC2/WG2 and UTC.
- Pluta, Olaf (1995). *Abbreviationes, the first electronic dictionary of medieval Latin abbreviations*.
- (2020). “Abbreviations.” In: *The Oxford Handbook of Latin Palaeography*. Ed. by Frank T. Coulson and Robert G. Babcock. Oxford: Oxford University Press, pp. 9–24.
- Robinson, Peter (n.d.). “Why Interfaces Do Not and Should Not Matter for Scholarly Digital Editions.” <https://www.slideshare.net/PeterRobinson10/why-interfaces-do-not-and-should-not-matter-for-scholarly-digital-editions>.
- Urbańczyk, Stanisław (1983). *Die altpolnischen Orthographien des 16. Jahrhunderts*. Ed. by Stanisław Urbańczyk and Reinhold Olesch. Vol. 37. Slavistische Forschungen. Köln-Wien: Böhlau.

# Graphemic Normalization of the Perso-Arabic Script






Raiomond Doctor, Alexander Gutkin,  
Cibu Johny, Brian Roark & Richard Sproat

*Abstract.* Since its original appearance in 1991, the Perso-Arabic script representation in Unicode has grown from 169 to over 440 atomic isolated characters spread over several code pages representing standard letters, various diacritics and punctuation for the original Arabic and numerous other regional orthographic traditions (Unicode Consortium, 2021). This paper documents the challenges that Perso-Arabic presents beyond the best-documented languages, such as Arabic and Persian, building on earlier work by the expert community (ICANN, 2011; 2015). We particularly focus on the situation in natural language processing (NLP), which is affected by multiple, often neglected, issues such as the use of visually ambiguous yet canonically nonequivalent letters and the mixing of letters from different orthographies. Among the contributing conflating factors are the lack of input methods, the instability of modern orthographies (e.g., Aazim, Mansour, and Pournader, 2009; Iyengar, 2018), insufficient literacy, and loss or lack of orthographic tradition (Jahani and Korn, 2013; Liljegren, 2018). We evaluate the effects of script normalization on eight languages from diverse language families in the Perso-Arabic script diaspora on machine translation and statistical language modeling tasks. Our results indicate statistically significant improvements in performance in most conditions for all the languages considered when normalization is applied. We argue that better understanding and representation of Perso-Arabic script variation within regional orthographic traditions, where those are present, is crucial for further progress of modern computational NLP techniques (Conneau et al., 2020; Muller, Anastasopoulos, Sagot, and Seddah, 2021; Ponti et al., 2019) especially for languages with a paucity of resources.

## 1. Introduction

The Modern Perso-Arabic script derives from the fourth century North Arabic script, which in turn was adapted from the Nabatean Aramaic

---

Raiomond Doctor  0009-0005-8684-4937, on contract from Optimum Solutions, Inc. Alexander Gutkin  0000-0001-6327-4824, Cibu Johny  0009-0008-4220-5414, Brian Roark  0000-0002-7292-6246 & Richard Sproat  0000-0002-9040-5196

Google Research: India, United Kingdom, United States, and Japan  
E-mail: {raiomond, agutkin, cibu, roark, rws}@google.com

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 315–376. <https://doi.org/10.36824/2022-graf-gutk>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

script to write the Arabic language (Bauer, 1996; Gruendler, 1993). Due to the spread of Islam throughout much of Africa, Asia, and parts of Europe, it has come, in its various adapted forms, to be one of the most widely used scripts in the modern world. Due to its reasonable flexibility in representing phonological structure, the script was adopted to write a large number of languages spanning diverse language families such as Afro-Asiatic, Indo-European, Niger-Congo, Turkic, and Sino-Tibetan, among others. Adaptations are found as far south as Southern Africa with Arabic having been used for Afrikaans (Kotzé, 2012) and Malagasy (Versteegh, 2001); as far east as East Asia for Chinese (Suutarinen, 2013) and Japanese (Kaye, 1996; Naim, 1971); and into Eastern Europe for writing languages of Muslim Slavs, such as Bosnian (Buljina, 2019). While many of these adaptations have not survived, the Arabic script and its derivatives are still used for scores of languages with a total population of speakers of over 600 million.<sup>1</sup> For some linguistic areas, such as the Dardic languages of Northern Pakistan, most of which were unwritten until very recently, the Perso-Arabic script is the *only* serious contender when developing a new writing system; see for example Torwali (Torwali, 2019),<sup>2</sup> and Palula (Liljegren, 2016).

While the original Semitic scripts were pure consonantal scripts (*abjads*), three letters—*alif* /ʔ/, *ya* /y/, and *waw* /w/—came to be used as *matres lectionis* to represent long vowels, and further diacritics were developed to (optionally) represent such features as short vowels and gemination (*shadda*), among others (Bauer, 1996).

The original North Arabic script was rather ambiguous, since Arabic had a larger consonant inventory than Aramaic, and some of the consonant letters had to do double duty—a problem exacerbated by the cursivization of the script. The resulting ambiguities were resolved by the use of various numbers of dots over or under the letters to disambiguate the various uses (Bauer, 1996; Kaplony, 2008), a system called *iʿjām* (إِعْجَام). For example the inferior dot in <ب> /b/ distinguishes it from <ن> /n/ with a single dot on top, and then again from <ت> /t/ with two dots on top, and again still from <ث> /θ/ with three on top. Though the set of consonants to be disambiguated is of course limited in Arabic itself, the *iʿjām*, once started, evolved into a productive way to produce new consonant symbols when the script was adapted to new languages. This has consequently allowed languages to have their “own” version of the Perso-Arabic script, where the only difference with the scripts used for a language’s neighbors is in the use of distinctive *iʿjām*-

1. [https://en.wikipedia.org/wiki/List\\_of\\_writing\\_systems/List\\_of\\_writing\\_systems\\_by\\_adoption](https://en.wikipedia.org/wiki/List_of_writing_systems/List_of_writing_systems_by_adoption)

2. <https://www.blog.google/around-the-globe/google-asia/torwali-language-and-its-new-android-keyboard/>

augmented consonants. This is true, for example, for adaptations of the script to the many Dardic languages, where each has one or two consonant symbols not found in the scripts of its neighbors.

As noted above, the Arabic script and its derivatives include diacritics that allow one to specify all vowels and other phonetic features such as gemination. However in the normal daily use of the script for Arabic and other languages, these are typically omitted. In Arabic this means that the script is still technically an *abjad*, since the written symbols mostly represent consonants. However to varying degrees, the derived scripts have departed from this, and some of them are full alphabets. Thus according to Kaye (1996), the Persian writing system is an *abjad* as are Urdu and Jawi, the old Malay Arabic-based writing system; however the Kurdish and Uyghur writing systems are alphabets. Parallel developments occurred with adaptations of the Hebrew *abjad* so that Yiddish orthography (Aronson, 1996) is an alphabet.

Historically each geographic region posed its own unique sociolinguistic challenges resulting in the emergence of different adaptation strategies and orthographic traditions across South Asia (Qutbuddin, 2007; Wink, 1991), Southeast Asia (Abdullah et al., 2020; Kratz, 2002; Ricci, 2011), and Africa (Mumin, 2014; Ngom and Kurfi, 2017), among other regions (Castilla, 2019; Suutarinen, 2013). This diversity as well as the flexible nature of the script is reflected in a large and growing inventory of Perso-Arabic code points in the Unicode standard (Unicode Consortium, 2021) with accompanying ambiguities associated with representing the script using the digital medium that we briefly outline in §2. We then provide an overview of some of the regional orthographies for eight languages selected from diverse language families in §3. A significant amount of digital representation ambiguities manifest by these orthographies is resolved computationally using finite-state normalization methods for Perso-Arabic, described in §4, that we developed for this purpose.<sup>3</sup> We study the effects of normalization of real-world text using statistical and neural techniques, and present our findings in §5.1 and §5.2, respectively. The code and the results accompanying the experiments have been released.<sup>4</sup>

## 2. Perso-Arabic in the Digital Medium

As was mentioned previously, an important feature which has led to the adoption of the script by different cultures to use it to transcribe their

---

3. <https://github.com/google-research/nisaba>

4. [https://github.com/google-research/google-research/tree/master/perso\\_arabic\\_norm](https://github.com/google-research/google-research/tree/master/perso_arabic_norm)

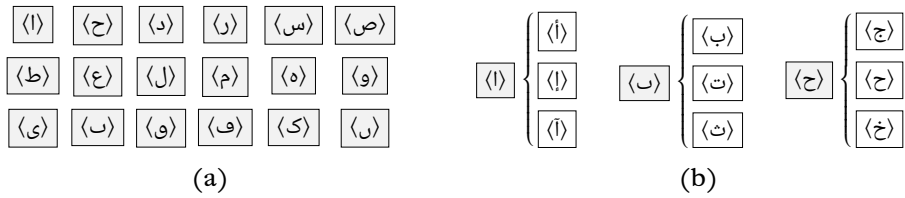


FIGURE 1. Core *rasm* shapes, or *archigraphemes* according to Milo (2002), of Arabic shown in (a), and examples of Arabic letters derived with *iḥjām* demonstrating their disambiguation function in (b), after (Nemeth, 2017)

language, is its very flexibility.<sup>5</sup> At its core the Arabic script comprises 18 basic shapes often referred to as *rasm* (رسم) or “drawing” (Daniels, 2013; Kurzon, 2013). These can be modified in various ways: apposing one to four dots (*iḥjām*) placed above, below or inside a character (as shown in Figure 1); using modifier signs such as the subscript or superscript *small hamza*; placing diacritics or *tasbkīl* (تشكيل) and in certain cases even adding a new shape based on the basic Arabic template. Thus for example Urdu, discussed in more detail in §3.1, substantially expanded the original Arabic writing system adapting it to its phonology by introducing additional *iḥjām* characters, modifiers, and even creating new shapes such as the *bari yeb* <ع> or the *heb do chashmee* <ه> for handling aspiration.

Similar to Brahmic scripts, the Perso-Arabic script often provides more than one way to compose a character in the digital medium (Unicode Consortium, 2021). For example, the *alef with madda above* letter can be composed in two ways: as a single character <آ> (U+0622) or by adjoining *madda above* to *alef* (U+0627 followed by U+0653). This results in presentation ambiguity and the Unicode standard provides a certain number of canonical normalization forms, such as the Normalization Form C, or NFC, to handle such cases (Whistler, 2021). A normalization process is required to convert strings to such canonical forms. In contrast to Brahmic script normalization, where atomic forms are normalized to their decomposition, Perso-Arabic normalization in NFC involves combining decomposed elements into a single glyph. Thus the individual *alef* and *madda above* will be normalized to a single glyph. Our investigations have found many cases of this kind of *visual ambiguity* in the Perso-Arabic script beyond what is covered in NFC.

Some of these visual ambiguities are illustrated by a simple example in Table 1, where six alternate representations for the Arabic word for “leader” are shown rendered in Naskh style along with the corre-

5. Scripts are sets of characters used jointly in written representation while writing systems additionally consist of the rules and conventions used when employing a script for a particular language.

TABLE 1. Six different spellings of the Arabic word for “president” (MSA: /ra.ʔi:s/) rendered in Naskh. For each row, the Unicode character differences with the Unicode string in the first row are highlighted. The last column indicates the type of transformation required to bring the relevant Unicode string to the canonical form displayed in the first row of the table.

| Display | C <sub>1</sub> | Unicode Character Sequence  |                    |                  |             | C <sub>5</sub> | Transformation        |
|---------|----------------|-----------------------------|--------------------|------------------|-------------|----------------|-----------------------|
| رئيس    | <i>reb</i>     | <i>yeb with hamza above</i> | <i>yeb</i>         | <i>seen</i>      |             |                |                       |
|         | U+0631         | U+0626                      | U+064A             | U+0633           |             |                |                       |
| رئيس    | <i>reb</i>     | <i>yeb</i>                  | <i>hamza above</i> | <i>yeb</i>       | <i>seen</i> |                |                       |
|         | U+0631         | U+064A                      | U+0654             | U+064A           | U+0633      |                | Unicode NFC           |
| رئيس    | <i>reb</i>     | <i>alef maksura</i>         | <i>hamza above</i> | <i>yeb</i>       | <i>seen</i> |                |                       |
|         | U+0631         | U+0649                      | U+0654             | U+064A           | U+0633      |                | Visual Normalization  |
| رئيس    | <i>reb</i>     | <i>yeb with hamza above</i> | <i>farsi yeb</i>   | <i>seen</i>      |             |                |                       |
|         | U+0631         | U+0626                      | U+06CC             | U+0633           |             |                | Visual Normalization  |
| رئيس    | <i>reb</i>     | <i>farsi yeb</i>            | <i>hamza above</i> | <i>yeb</i>       | <i>seen</i> |                |                       |
|         | U+0631         | U+06CC                      | U+0654             | U+064A           | U+0633      |                | Reading Normalization |
| رئيس    | <i>reb</i>     | <i>farsi yeb</i>            | <i>hamza above</i> | <i>farsi yeb</i> | <i>seen</i> |                |                       |
|         | U+0631         | U+06CC                      | U+0654             | U+06CC           | U+0633      |                | Reading Normalization |

sponding Unicode sequences ranging from four to five characters. The spelling in the first row of the table is the correct spelling of the word in Modern Standard Arabic (MSA) orthography. The visual forms in the second, third and fourth rows are visually identical to the correct spelling, but are represented digitally as distinct sequences of characters. The second example, while ambiguous, is handled by the Unicode NFC normalization, which brings it to the canonical form provided by the first row by rewriting a (decomposed) sequence of *yeb* and *hamza above* as its canonical single-letter counterpart *yeb with hamza above*. The form in the third row is more problematic. It arises from a five-character sequence which has *alef maksura* instead of *yeb* and is also visually identical to words in rows one and two. However, unlike the spelling in the second row, the sequence *alef maksura* followed by *hamza above* does not have a canonical composed form under Unicode.<sup>6</sup> Hence, while this

6. See the discussion in (Pournader, 2010) on how this came about.

form is visually identical to the first and second forms, it is treated as a distinct word in the digital medium. Similarly, the example in the fourth row is visually identical to the preceding examples, but arises due to using *farsi yeb* instead of *yeb* which is illegal under MSA orthography. No standard transformation is provided by the Unicode to cope with cases like this example as they are orthography- and language-specific. We refer to the class of normalizations that result in something that is visually identical as *visual* normalization (Johny, Wolf-Sonkin, Gutkin, and Roark, 2021).

The second group of ambiguities is illustrated by the last two rows of Table 1. The fifth and sixth examples, while visually identical to each other, differ slightly from the reference in extra *ījām* dots produced by combining *farsi yeb* with *hamza above* word-internally. Similar to the examples in rows three and four, no fallback normalization strategy is provided by the Unicode standard to handle such cases as it is not clear without prior context which orthography is intended. We refer to the class of normalizations that result in something that is not visually identical as *reading* normalization (Gutkin, Johny, Doctor, Wolf-Sonkin, et al., 2022).

As we mentioned above, unlike the canonical Unicode transformations, it is impossible to define most of visual and reading normalizations outside the orthographic context. Consider a sequence consisting of *waw* (U+0648) followed by *damma* (U+064F) whose visual form <ّ> is identical to letter *u* (U+06C7) used in Kazakh, Uzbek and Uyghur Perso-Arabic orthographies among some others (Aqtay, 2020; Haralambous, 2021). Normalization of *waw* and *damma* to its “canonical” form *u* should only be performed for these languages. We introduce the visual and reading script normalization framework more formally in §4.

One could argue that the ambiguities described above are not problematic and are the natural outcome of the specific properties of the script (e.g., its cursive form and the presence of positional variants), the vast number of orthographic adaptations and specifics of its implementation in digital medium. However, as we show in §5, the resolution of such ambiguities on a language-by-language basis positively impacts the quality of computational models of natural language. Furthermore, the visual ambiguities manifest by various Perso-Arabic writing systems represent a growing challenge to cybersecurity. From the standards’ perspective this is being gradually addressed by the Internet Corporation for Assigned Names and Numbers (ICANN) by developing a robust set of standards for representing various Internet entities in Perso-Arabic script, such as domain names, with particular focus on visually confusable character variants (ICANN, 2011). In addition, security implications, such as development of mechanisms for protection against *phishing* and *spoofing* attacks, are actively studied by the relevant



cybersecurity literature on Perso-Arabic (Ahmad and Erdodi, 2021; El-sayed and Shosha, 2018; Hussain et al., 2016).

A detailed analysis of the causes for the various types of Perso-Arabic script ambiguities described above are outside the scope of this work. It is however worthwhile to briefly mention some of them here. One of the causes is the relative complexity and several special properties of the Perso-Arabic script itself. The script has four key orthographic properties that are relevant here: (i) relative visual similarity of the *rasm* forms; (ii) *allography*, i.e., positional variants of letters (initial, medial, final and isolated); (iii) cursivity/ligaturing, and (iv) *non-linearity*, the extensive and sophisticated use of various types of *tashkīl* and *iqjām* (Yassin, Share, and Shalhoub-Awwad, 2020). The combination of all these properties was demonstrated to lead to relatively more involved visual processing of the script (compared to languages that use the Latin script) in the psycholinguistics and neuropsychology literature on reading for MSA (Boudelaa, Perea, and Carreiras, 2020; Eviatar and Ibrahim, 2014; Hermena and Reichle, 2020), but also for the Uyghur Perso-Arabic orthography (Yakup, Abliz, Sereno, and Perea, 2015).

The Perso-Arabic script support in Unicode is ever-growing, which is reflected by the number of recent proposals for new characters to better reflect the existing low-resource orthographies (Chitralli, 2020a,b; Evans and Warren-Rothlin, 2018; Patel, Riley, and MacLean, 2019) as well as to ease the encoding of the Quranic literature (Sh., 2022). The process of updating the standard is usually time-consuming, as demonstrated by the case of Torwali, which took two years from the time of the original proposal (Bashir, Hussain, and Anderson, 2006) to encode the missing letter *bab with small arabic letter tab above* <ځ> (U+0772). This letter completed the full character inventory for this emerging orthography in Unicode, which facilitated further developments of linguistic resources (Uddin and Uddin, 2019). As we found in our experiments, in the absence of the required characters, visually confusable variants or sequences of variants from foreign orthographies are often used by the Unicode-compliant input methods and converters from non-Unicode compliant fonts. Conversely, these methods take time to catch up with the Unicode standard once it introduces the missing features. To this one can add multiple confounding factors involved in the modern evolution of orthographies for hitherto unrecorded languages, which leads to rich orthographic diversity even among neighboring languages. For example, according to Bashir (2015, p. 14), the retroflex voiceless sibilant /ʂ/ present in several languages of northwestern Pakistan is represented differently by the regional writing systems: they all share the same *rasm* shape for letter *seen* <س> modified as: (1) *seen with small arabic letter tab and two dots* <ڄ> (U+0770) for Khowar; (2) *seen with extended arabic-indic digit four above* <س̣> (U+077D) for Burushaski; (3) *seen with four dots above* <س̣̣̣̣> (U+075C) for Torwali; (4) *seen with two dots vertically above* <س̣̣> (U+076D) for

Gowri; and (5) the corresponding two distinct characters for Kalasha and Shina remain unrepresented in Unicode.

### 3. Perso-Arabic Script Diaspora: Selected Language Summaries

In what follows, we briefly delve into some adaptations of Perso-Arabic script. We limit our discussion to Perso-Arabic orthographies of eight languages, some of which are written in several scripts. The five languages from Indo-European family are Central Kurdish (Sorani), Kashmiri, Punjabi, Sindhi and Urdu. Two further languages, South Azerbaijani and Uyghur, come from the Turkic family. Finally, we provide a brief overview of the Perso-Arabic orthography called Jawi for Malay from the Austronesian family. The concise language-specific letter inventories are provided in §A. Our software covers more orthographies, such as Balochi, Dari, Modern Standard Arabic, Pashto, Persian and Uzbek, yet we felt that our choice of the above eight writing systems is representative of the kinds of normalization challenges one is likely to encounter.

#### 3.1. Urdu

Ethnologue classes Urdu as the tenth most spoken language in the world with over 70 million speakers using Urdu as their first language.<sup>7</sup> The national language of Pakistan, one of the 22 official languages of India, and a registered dialect in Nepal, Urdu is also spoken and used in 30 odd countries.<sup>8</sup>

The origins of Urdu are debatable and some scholars trace it back to the 6th century CE (Schmidt, 2007) but it was the Muslim invasion of Sindh in 711 which acted as a catalyst. By the time of the Mughal Empire and at the end of the 18th century it was the lingua franca around Delhi and was called *Zaban-e-Urdu*,<sup>9</sup> the word Urdu derived from the Turkic word *ordu* for “army” (Lelyveld, 1994).<sup>10</sup> The expansion of the Sultanate to the south gave rise to Dakhani Urdu (Mohamed, 1968). Urdu has a close association with Hindi since they share a common Indo-Aryan origin. Whereas Urdu is written in Perso-Arabic, Hindi uses Devanagari. The difference is best seen in the two versions of “The Chess Players”

---

7. <https://www.ethnologue.com/guides/ethnologue200>

8. <https://www.ethnologue.com/language/urd>

9. Urdu was also called *Hindi*, *Hindustani*, *Dehlavi*, and *Lasbkari*. But the term Urdu became most acceptable.

10. <https://www.rekhtadictionary.com/meaning-of-urdu>

story written by Munshi Premchand, who authored the same story in both languages (Davis, 2015).<sup>11</sup>

TABLE 2. Urdu in Nastaliq (top), Naskh (center) and the corresponding transliteration (bottom). Samples taken from a poem (*nazm*) by Faiz Ahmed Faiz

| Nastaliq   |
|--|
| تیرا غم ہے تو غم دہر کا جھگڑا کیا ہے<br>تیری صورت سے ہے عالم میں بہاروں کو ثبات                |
| Naskh  |
| تیرا غم ہے تو غم دہر کا جھگڑا کیا ہے<br>تیری صورت سے ہے عالم میں بہاروں کو ثبات                |
| Transcription  |
| terā ḡham hai to ḡham-e-dahr kā jhagḏā kyā hai<br>terī sūrat se hai aalam meñ bahāroñ ko sabāt |

Urdu as used in India and Pakistan is written in the Nastaliq style—a writing style developed in Iran from the Naskh style around the 13th century. Easy to write by hand, it posed problems when ported to metal type. Digital typography has to a certain extent solved the problem and text can be seen in the Nastaliq style, however media on the Web prefers Naskh (Parhami, 2020). A sample of two lines of a *nazm* by Faiz Ahmed Faiz in Table 2 demonstrates the differences between the two styles.<sup>12</sup>

The Urdu writing system is an *abjad*, borrowed from Persian which in turn is borrowed from Arabic. Persian added four characters (<پ>, <ز>, <گ>, and <چ>) to the 28 basic characters borrowed from Arabic, bringing the total to 32. Persian further modified the character set by replacing the Arabic characters <ي> and <ك> with <ی> and <ک>, respectively. To these 32 Persian characters, Urdu added: (1) the three letters <ٹ>, <ڈ>, and <ظ> to accommodate retroflexes; (2) <ں> to handle nasalization; (3) the *two-eyed be* <ھ> to accommodate the aspirated forms of 17 or 18 letters; (4) the *yeh baree* <ے> to represent /e/ at the end of the words; and (5) <و> *gol be*, also called *choṭī be*. Since diacritics are a systematic component of Perso-Arabic, this was possible without upsetting the graphic equilibrium of the script (Coulmas, 1999, p.560). The added *high hamza* placed above *farsi yeh*, *yeh baree*, *beb goal* and *waw* is used to create additional values and the *teh marbuta* <ة> marks feminine gender for

11. <https://thewire.in/culture/why-the-perso-arabic-script-remains-crucial-for-urdu>

12. <https://www.rekhta.org/nazms/>

nouns and adjectives. A further 5 characters were added to represent the 10 vowel phonemes, and an additional 5 to 10 diacritics were used when precision was needed.<sup>13</sup>

### 3.2. Punjabi (Shahmukhi)

TABLE 3. Shahmukhi in Nastaliq (top), Naskh (center) and the corresponding transliteration (bottom). Samples taken from a poem by Baba Farid

| Nastaliq   |
|--|
| کافی آوو سکھی سہلیو مل مسلت گوئے<br>آپو آپنی گل نوں بھر ہنجھو روئے<br>کھڈے لالچ لگیاں میں عمر گوائی<br>کدے نہ پونی ہتھ لے اک تندڑی پائی                                |
| Naskh  |
| کافی آوو سکھی سہلیو مل مسلت گوئے<br>آپو آپنی گل نوں بھر ہنجھو روئے<br>کھڈے لالچ لگیاں میں عمر گوائی<br>کدے نہ پونی ہتھ لے اک تندڑی پائی                                |
| Transcription  |
| kaafi—aavo sakhi saheliyo mil maslat goiye<br>aapo aapni gal nuun bhar hanjhū roīye<br>khaDe lālach lagyañ maiñ umar gavā.ī<br>kade na puuni hath lae ik tandaḌī paa.ī |

The Shahmukhi writing system is used to record Punjabi in the Perso-Arabic script. According to Ethnologue this language is mainly spoken in Pakistan but also in other countries, especially in Punjab in India, with a total number of Punjabi speakers around 66 million.<sup>14</sup> The language is also known as Jangli, Lahanda, Lahnda, Lahndi, Panjabi, Panjabi Proper and Punjabi. Historically Shahmukhi was used by Sufi poets of the then Punjab region. One of the earliest instances of this writing system is its use by the Sufi poet of Punjab, Baba Farid in the 12th century (Singh and Gaur, 2009). After the partition of India, Shahmukhi became the writing system of choice for writing Punjabi by the Muslim population in Punjab. Hindus and Sikhs in the Indian state of

13. The number of characters is debated, cf. <https://www.dawn.com/news/919270>.

14. <https://www.ethnologue.com/language/pnb>

Punjab adopted the Brahmic Gurmukhi script to write Punjabi, giving rise to Eastern Punjabi (Grewal, 2004). The relationship between Shahmukhi and Gurmukhi closely parallels that between Urdu and Hindi. Shahmukhi is an *abjad* and is written from right to left. It was highly influenced by Persian, but the present day writing system was modified to suit the requirements of the Punjabi language and as in the case of Urdu, a considerable number of characters were added. Like Urdu, Shahmukhi favours Nastaliq, but Naskh is used by digital media on the web. A sample of four verses from a poem by Baba Farid in both styles is shown in Table 3.

Urdu and Shahmukhi share the same character set, except that Shahmukhi admits a few more letters. The number of characters in Shahmukhi, like Urdu, is a matter of debate and some scholars admit four more characters <ٲ>, <ٲ>, <ٲ>, and <ٲ> in addition to the retroflex lateral *lla* <ٲ> and the retroflex nasal *nnā* <ٲ> (Bashir and Connors, 2019, pp.62, 77). Of these, our analysis shows that only <ٲ> and <ٲ> seem to be in use. The <ٲ> character is used to mark end-of-word nasals. Like Urdu, the *two-eyed he* <ٲ> is used to accommodate the aspirated forms of 17 or 18 letters. To these can be added the *high bamza* (U+0674) placed above *farsi yeh*, *yeh baree*, *heh goal* and *waw* to create additional values. Finally, five characters are used to represent the 10 vowel phonemes, and an additional 5 to 10 diacritics are used when precision is needed in cases such as consonant clusters or gemination. Shahmukhi and Urdu are thus mutually intelligible as writing systems.

### 3.3. Sindhi

Sindhi is an Indo-Aryan language spoken by the inhabitants of Sindh in the western part of the Indian subcontinent. It is one of the official languages of Pakistan and one of the 22 scheduled languages in India. Thanks to the Sindhi Diaspora it is spoken in quite a few countries and as per Ethnologue, has over 33 million speakers around the world.<sup>15</sup> Sindhi is recorded both in Perso-Arabic as well as Devanagari scripts. The traveler Al-Biruni in his *Tarikh-al-Hind* states that Sindhi was written in three scripts: Ardhanagari, Mahajani and Khudabadi (Sachau, 1910). But, the standardization of the Sindhi Perso-Arabic writing system (“arabi Sindhi”) dates back to the 19th century. Prior to that, Sindhi Muslims had made attempts to write the language using Arabic, but the formal character set of Sindhi, as it is known today, goes back to 1853 when it was standardized by the British colonial authorities (Dow, 1976) and a set of 52 letters to accommodate the complexities of the sound system of Sindhi was identified. Sindhi is an *abjad* but unlike Urdu or Shah-

15. <https://www.ethnologue.com/language/snd>

TABLE 4. Sample of Sindhi in Naskh (top), Devanagari (center), and the corresponding Latin transliteration (bottom). Samples taken from a poem by Shah Abdul Latif Bhittai

| Naskh  |
|--|
| سڄڻ ۽ سائيهه ڪنهن اناسي وسري،<br>حيث تينن کي هوءَ، وطن چن وساري.                 |
| Devanagari   |
| सज्जन ऐं साणेहु कं हिं अणासी विसरी,<br>हैफ तनीं खे होइ, वतनु जिनि विसारी.        |
| Transliteration  |
| sajanu ain sanehu kanhin anasi visri,<br>haif tanin khe hoi, vatanu jini visari. |

mukhi, Sindhi is only written using the Naskh style. A sample stanza from “Shah Jo Risalo” by Shah Abdul Latif Bhittai (Lajwani and Mirjat, 2021) shown in Table 4.

The addition of digraphs and the *hamza* over *yeh* and *waw*, as well as the diacritics to indicate the short vowels placed above *alef* and *waw*, brings the size of modern Sindhi letter inventory to 64 (Lekhwani and Lekhwani, 2014). For short vowels in particular, the following four letters <ا>, <اِ>, <اُ>, <ؤ>, composed by placing diacritic marks *fatha*, *kasra* and *damma* over *alef*, and the *damma* over *waw*, were added. To accommodate the large number of characters in its repertoire, Sindhi modified the Arabic *rasm* by addition of more *ijām* dots.

Certain features of the character set of Sindhi make for the uniqueness of the writing system. Unlike Urdu or Shahmukhi, the *high hamza* is already accommodated over <ؤ> and <ئ>. Sindhi admits four implosives <ڳ> *ga*, <ڄ> *ja*, <ڏ> *da*, <ڀ> *ba*, and two single letter words <ء> *ain* (“and”), and <م> *men* (“in”). Like Urdu, Sindhi has four letters to indicate /z/: <ض>, <ظ>, <ز>, <ذ>; three letters for /s/: <س>, <ص>, <ت>; and two letters for /h/: <ح> and <ھ>. However, unlike Urdu which uses the *two eyed he* or *be do chashmi* <ھ> to mark the aspirates, Sindhi has individual characters for all the aspirates with the exception of <ڱ> *gha*, <ڄھ> *jba*, and <ڙھ> *rba*. Vowel diacritics are not normally used. However if needed Sindhi has three diacritics used to indicate the short vowels. Additional diacritics are used to mark consonant clusters (*sukun*, U+0652) and gemination (*shadda*, U+0651).

TABLE 5. A sample of Kashmiri proverbs

| Nastaliq         | Transliteration      | Translation                                      |
|------------------|----------------------|--|
| کنو کن بتر لادن  | Kanav kin batı ladun | Stuff rice through the ears: to overfeed         |
| کنس کنس بتر لادن | Kanas batı ladun     | Stuff the ear with rice: advice wasted on a fool |

### 3.4. Kashmiri

Kashmiri is a language from the Dardic family spoken in the Union Territory of Jammu and Kashmir, Himachal Pradesh and its outlying regions (Koul and Wali, 2015). Ethnologue lists around 7 million Kashmiri speakers in India and other countries.<sup>16</sup> It is a statutory language of provincial identity in Jammu and Kashmir<sup>17</sup> and is one of the 22 scheduled languages in India.<sup>18</sup> According to B. B. Kachru (2016), Kashmiri is the only Dardic language with a literary tradition and for which the written records have survived. Kashmiri is one of the three scheduled languages of India that is written using a Perso-Arabic script, the other two being Urdu and Sindhi. As in the case of Sindhi, Kashmiri is also written in Devanagari script, suitably modified to accommodate the sounds of the language.<sup>19</sup>

Historically Kashmiri was written in the Sharada script, an *abugida* from the Brahmic family (Khaw, 2015). Sharada fell into disuse because it could not represent the complex sound system of the language. Successive invasions of the region slowly led to the adoption of the Arabic script. By the 14th century, Muslim rule in Kashmir was established and Kashmiri in Perso-Arabic script was adopted (Yatoo, 2012). The writing system evolved with time and the Arabic *rasm* were suitably adapted to add new characters to the repertoire. Today the Perso-Arabic script is recognized as the official writing system for the language. It is written in both Naskh and Nastaliq; and although the latter is favoured as the desired style, digital media prefers Naskh owing to the non-availability of a Nastaliq font for the script. Kashmiri is renowned for its proverbs (R. Kachru, 2021; Koul, 2006) and a sample of two proverbs<sup>20</sup> in Nastaliq, with the corresponding transliterations and English translations, is shown in Table 5.

16. <https://www.ethnologue.com/language/kas>

17. In 2020, the Parliament of India passed a bill to make Kashmiri an official language of Jammu and Kashmir along with Dogri, Hindi, Urdu and English.

18. <https://rajbhasha.gov.in/en/languages-included-eighth-schedule-indian-constitution>

19. कौशुर (Koshur)

20. <https://kashmiridictionary.org/kanas-bati-ladun/>

Kashmiri is an *abjad*, but because all vowel sounds are regularly indicated in its orthography, the writing system is somewhat closer to an alphabet similar to Sorani Kurdish and Uyghur. The consonant inventory of Kashmiri consists of 37 letters. Some of these letters are shared with Urdu, Khowar and Shahmukhi orthographies, like the letter *rreh* <ڙ> for representing voiced retroflex flap /ɽ/ or the *ddal* <ڊ> for the voiced retroflex plosive /ɖ/. Kashmiri, like Urdu and Shahmukhi, uses the *two-eyed he* <ه> in the construction of aspirated consonants. However, unlike Urdu and Shahmukhi only six such digraphs are permitted: <هت>, <هڄ>, <هپ>, <هچ>, <هڙ>, and <هڪ>. The character <ي>, a *yeh with a ring below* needs special mentioning. Kashmiri uses <ي> to mark palatalization which is a common feature in the language.

Kashmiri has one of the largest inventories of vowel letters, which are arranged in eight pairs of short and long vowels.<sup>21</sup> Kashmiri uses the *kasra*, *damma* and *fatha* as short vowel diacritic markers. Kashmiri modifies the *waw* to add new vocalic values: <و> *waw with ring* to represent the sound /ɔ/; <ۆ> *waw with inverted v on top* for /o/; and <ؤ> *waw with inverted damma* for a long /u:/. Additionally, *yeh baree* with an *inverted small v* marks the short /e/. Two combining marks are unique to Kashmiri, these are the *wavy hamza above* and *wavy hamza below*. The first is always used in conjunction with *alef* and represents a long schwa /ə:/, while the second is used along with *alef* to represent /i:/.

Similar to Urdu and Shahmukhi, Kashmiri nasalisation is marked by the *noon ghunna* <ٲ> which can only occur in final position. When in medial position it is replaced by the letter *noon* <ن>. Kashmiri also uses two other combining marks to mark gemination using *shadda*, and *sukun/jazm* (also called a *vowel killer*) to mark consonant clusters. The rendering of standard *sukun* diacritic (U+0652) is unique to Kashmiri writing system and has the shape of an inverted <v>.

### 3.5. Central Kurdish (Sorani)

Sorani is the Perso-Arabic writing system used to write the Kurdish language in Iraq, mainly in Iraqi Kurdistan (Haig, 2018). This Indo-Iranian language is also spoken in regions adjoining Iran and Turkey. Ethnologue identifies three geo-linguistic variants of the language depending on where it is used: Central (Zimane Sorani), Northern (Kurmancî) and Southern (Kurdi Xwarîn or Pehlewani). The name Sorani derives from the Soran Emirate, located in the area known today as Iraqi Kurdistan. Ethnologue lists around 4.7 million Sorani speakers in Iraq and total

21. According to <https://r12a.github.io/scripts/arabic/ks.html>. Slightly different inventory is provided in <https://kashmiridictionary.org/category/learn-kashmiri/vowels-learn-kashmiri/>.



TABLE 6. Sorani sentence in Naskh and Latin translating as “Twenty million people are asking for tickets to participate in the last Led Zeppelin concert”

| Naskh   |
|---|
| <p>بیست ملیۆن کەس داواى بلیت دەکەن بۆ بەشداری کردن<br/>له دوا کۆنسێرتی لێد زێپالین</p>          |
| Latin   |
| <p>bîst miliyon kes dawayi bilît deken bo beşdarî kardîn<br/>lah dawa konisêrtî lêd zêpalîn</p> |

number of Sorani speakers in all countries at 5.3 million.<sup>22</sup> Unlike other languages in this study, Kurdish is not recognized today as the official language in any of the regions where it is used,<sup>23</sup> despite popular movements by Iraqi Kurds to give the language an official status.

Sorani in Perso-Arabic traces its origins to the Sulaimani region. The first trace of this writing system is found in “Mahdîname” by Mullah Muhammad Ibn ul-Haj completed circa 1762 (Bozarslan, Gunes, and Yadirgi, 2021). The rise of the Baban dynasty encouraged the growth of Sorani and it became a medium for prose and poetry (Khalid, 2015). This continued until the Baban dynasty was overthrown around 1856. However, under British rule in the 19th century, Sorani literature and journalism flourished and multiple attempts were made to standardize the writing system, which led to eventual codification of the Sorani alphabet in the 1920s (Campbell, 1994).

Unlike most other Perso-Arabic writing systems, Sorani is a true alphabet, the vowels being explicitly marked. Sorani is written in Naskh style. A sample text from the Pewan corpus (Esmaili et al., 2013) is provided in Table 6.<sup>24</sup> As is the case with all languages adopting Perso-Arabic, in order to represent the phonemic features of the language, Sorani has evolved a system of letters some of which are unique to this writing system. These include the three unique consonant letters that are constructed by adding dots <ف> for /v/ or appending a small *v* below or above: <ڤ> for /r/, and <ڭ> for /ɣ/. Similar to other Perso-Arabic writing systems, the vowel set borrows from the consonant set in that *yeb* and *waw* double as vowels and consonants. *Alef* is used as a vowel. The long /u:/ is indicated by doubling the *waw*. *Waw* and *yeb* with a small *v* above indicate /o/ and /e/, respectively.

Although the *kasra* is not part of the modern Sorani orthography, it is rarely used in some dictionaries for disambiguating certain pronuncia-

22. <https://www.ethnologue.com/language/ckb>

23. In 2006, Duhok Governorate began using Kurmanji as their official language as a way of resisting Sorani.

24. <https://sinaahmadi.github.io/resources/pewan.html>

tions,<sup>25</sup> where it is used to mark a short /I/ that is otherwise unrepresented in modern Perso-Arabic orthography.<sup>26</sup>

### 3.6. Uyghur

Uyghur, also written as Uighur, is a Turkic language spoken in the region in and around what is known as the Xinjiang Uyghur Autonomous Region in Northwest China.<sup>27</sup> Due to a politically and culturally motivated diaspora, Uyghur is spoken in Turkic countries such as Kazakhstan, Kyrgyzstan, Uzbekistan and Turkey, but also by the smaller Uyghur migrant communities elsewhere (Dillon, 2009). Ethnologue estimates the number of Uyghur speakers at around 10 million in China and total speakers in all countries at around 10.4 million. The modern Uyghur writing system should not be confused with Old Uyghur which was written using Sogdian script (Wilkins, 2016). Historically the writing system dates back to the 10th century when the Perso-Arabic script was introduced along with the spread of Islam and which evolved after considerable changes over the centuries into what is recognized as the modern Uyghur Perso-Arabic orthography (Brose, 2017). The writing system for the language underwent extensive changes, including being changed to the Cyrillic and Latin scripts and even the Pinyin romanization system for political reasons. It was not until 1982 that the Arabic Uyghur alphabet was reinstated (Dwyer, 2005). As of today the language has four writing systems: Uyghur Arabic used in the Xinjiang province of China, Uyghur Cyrillic in Kazakhstan, Uyghur Latin in Turkey and Uyghur Pinyin, which is not used much (Hamut and Joniak-Lüthi, 2015).

Unlike most other writing systems using Perso-Arabic, but like Sorani, Uyghur writing system is an alphabet, i.e., the vowels are explicitly marked. Uyghur is written in Naskh style, a sample of which is shown in Table 7.<sup>28</sup> As is the case with all languages which have adopted the Perso-Arabic script, in order to represent the phonemic features of the language, Uyghur writing system has evolved an original repertoire of letters. Apart from the letters borrowed from the original Arabic script, four letters are derived from Persian writing system: <پ>, <چ>, <ژ>,

25. Private correspondence from Aso Mahmudi (2022).

26. According to Ahmadi (2019, §2.2, p.3), the corresponding letter of Latin-based orthography of Kurmanji dialect is ⟨i⟩.

27. <https://www.ethnologue.com/language/uig>

28. The Latin text obtained from UygurAvazi newspaper (<https://uyguravazi.kazgazeta.kz/>) was converted to Uyghur through a script converter from <http://www.elipbe.com>.

TABLE 7. Uyghur sample in Naskh transliterated from Latin text

| Naskh  |
|--|
| پره زېدەنت ماراسىم قاتناشقۇچىلىرىنى ۋاتان قىمايچىسى كۈنى<br>ۋا غالىبىيات كۈنى بىلەن تاپرىكلاپ قارىيى كەزىماتچىلارنىڭ<br>مىنلاشكا ئالاقىدا قاسسا قوشۇۋاتقانلىغىنى ئاتاپ كورساتتىن           |
| Latin  |
| prezident marasim qatnashquchilirini vatan qimayichisi kuni<br>va ghalibiyat kuni bilan tabriklap qarbiy khizmatchilarninh<br>taminlashka alaqlida qassa qoshuvatqanlighini atap korsattin |

TABLE 8. A line from South Azerbaijani Wikipedia: “Mirza Shafi Vazeh—Azerbaijani poet, thinker, enlightener and teacher”

| Nastaliq  |
|---|
| میرزا شفیع واضح آذربایجان شاعیری و موثقیری، معارفچی و پداقوق. |
| Naskh   |
| میرزا شفیع واضح آذربایجان شاعیری و موثقیری، معارفچی و پداقوق. |

<گ>; the <ك>, which represents a velar nasal, common to Turkic languages, is derived from the Arabic *kaf* <ك> with three dots positioned above the letter. The *two-eyed he* <ه> is also used, similar to Kazakh, Urdu, Sindhi and Shahmukhi among other languages.

Extensive use of the *waw* is made, which is modified in productive ways to represent the vowels: <ؤ> with a superscript *alef* to represent the sound /yu/, <ۇ> with a *small v* on top to represent a front rounded vowel /ø/, <ۇ> with a *damma* on top for a long /u:/ and <ۇ> with *three dots* on top represents the semivowel /w/. Uyghur uses the Arabic *yeb* <ي> for the semivowel /j/, the *alef maqsura* <ى> for the /i/ and <ې> *yeb* with two dots below to represent /e/. Additional combining marks are used to mark consonant clusters (Arabic *sukun*, U+0652) or gemination (Arabic *shadda*, U+0651).

### 3.7. Southern Azerbaijani

Azerbaijani, also known as Azeri, Azari, Azeri Turkish and Azerbaijani Turkish, belongs to the Turkic language family, more specifically to the Western Oghuz branch (Mokari and Werner, 2017). It is spoken by over 23 million people, mainly in Azerbaijan, Iran, Georgia, Rus-

sia and Turkey, and also in Iraq, Syria and Turkmenistan.<sup>29</sup> Two varieties of the language are recognized: Northern Azerbaijani and Southern Azerbaijani. Northern Azerbaijani is spoken in the Republic of Azerbaijan, where it is the official language. Southern Azerbaijani spoken by around 14.6 million people is confined to the northwest of Iran and is often called *Turki* (تورکی).<sup>30</sup> Due to migrations and trade it is also used in parts of Iraq and Turkey, and in Afghanistan and Syria. Whereas Northern Azerbaijani uses either the Cyrillic script (in Dagestan) or Latin (the official script in Azerbaijan), Southern Azerbaijani uses the Perso-Arabic script. The Naskh style is favoured in day to day use but Nastaliq is sometimes used, mainly for book titles and also for handwriting, as demonstrated in Table 8.

Historically Old Azeri (*Ādari*) was the Indo-Iranian language spoken in Persian Azerbaijan before the arrival of the Turkic-speaking populations to the region (Yarshater, 2011). The language was gradually replaced with Turkish as the migration of Turkic speakers increased and by the 18th century Turkish was recognized as the language of Azerbaijan, although the name *Ādari* was retained as Azeri and traces of Old Azeri can still be found in Turkish today (Bosworth, 2011). The arrival of the muslim Turkish speakers in South Azerbaijan was accompanied by the Perso-Arabic *abjad* which became the official script of Azerbaijan until the 1920's, when, for political reasons, competing Cyrillic and Latin scripts entered the scene (Hatcher, 2008). The Azerbaijani Perso-Arabic writing system saw considerable mutation over the centuries: 28 letters (all from Arabic) initially, increased to 32 letters with additions from Persian and, finally, 33 letters due to an addition from the Ottoman Turkish. None of these solutions were found suitable for Azerbaijani and reforms were proposed during the 19th and 20th centuries which finally created the character set of Southern Azerbaijani as it is known today.<sup>31</sup>

The modern inventory consists of 42 letters. The majority of letters are borrowed from the Arabic and Persian orthographies. Nine letters (<ذ>, <ژ>, <ص>, <ض>, <ط>, <ظ>, <ع>, <ح>, and <ث>) are exclusively used for spelling Persian and Arabic loanwords and names. An extra letter *kebeh with three dots above* <ک> is used to indicate the voiced velar nasal /ŋ/, similar to Uyghur (Daniels, 2014, p. 31). Like all Turkic languages, Azerbaijani has a rich vowel system (Johanson and Csató, 2021). Three core shapes <ا>, <و>, and <ی>, modified with various diacritics, form the letters of the vowel set. Letter <ئ> represents the sound /e/, <ئ> the unrounded back vowel /u/ and <ي> represents /i:/. The *rasm* for *waw* is adapted in four ways. Apart from the intrinsic value of

29. <https://www.ethnologue.com/language/aze>

30. <https://www.ethnologue.com/language/azb>

31. For example, in modern Azerbaijani, letter *kebeh* <ک> has replaced the older Arabic *kaf* <ک>.

TABLE 9. Sample of Jawi in Naskh (top) and Rumi (bottom)

| Naskh  |
|--|
| ينله كراغن سواتو ماده مغارغن شعير نرلاو اينده،<br>ممبتولي جالن تمقت برقيبنده،<br>د سانله ايتيكات دقريتولي سوده                             |
| Rumi   |
| inilah gerangan suatu madah mengarangkan syair terlalu indah,<br>membetuli jalan tempat berpindah,<br>di sanalah i'tikat diperbetuli sudah |

*waw*, <ؤ> is used for /u:/, <ؤ> for the front open rounded vowel /y/, <ؤ> for the front rounded vowel /ø/ and the digraph *waw with sukun* <ؤ> represents /o/. The letter <ه> is used only in the final form to mark the diphthong /ae/. In addition, Azerbaijani orthography admits three combining marks, *fatha*, *damma* and *kasra*, to mark short vowels, and also additional diacritics to mark consonant clusters (*jazm/sukun*, U+0652) or gemination (*shadda*, U+0651).

### 3.8. Malay (Jawi)

Jawi is a Perso-Arabic writing system used for recording the Malay language from the Austronesian family and several other languages of Southeast Asia (Kratz, 2002). With the advent of Islam in Southeast Asia around the 14th century, the Pallava script, Nagari, and old Sumatran scripts which were used in writing Malay, were replaced by the Perso-Arabic script and by the 15th century Jawi had spread to Brunei, Indonesia and even Thailand due to trading (Coluzzi, 2020). Its dominance remained till the 20th century when Jawi was replaced by the Latin script (*Rumi*) and was confined to religious and cultural rituals. Today apart from Malaysia, Jawi has the status of an official writing system in Brunei and also in Indonesia, where Jawi has been assigned a regional status (Abdullah et al., 2020). Unlike Urdu or Shahmukhi, and, to a lesser extent Persian, Jawi favours the Naskh style, demonstrated by the sample quatrain from “Syair Perahu,” a Sumatran Sufi poem (Braginsky, 1975), in Table 9.

In addition to the 28 basic characters from Arabic,<sup>32</sup> Jawi added extra characters to suit its requirements and introduced the following: <چ> *ca*,

32. Some scholars, like R. O. Windstedt, believe that Jawi borrowed the characters from the Persian, rather than Arabic, orthography (Windstedt, 1961).

<ڠ> *nga*, <ڤ> *pa/fa*,<sup>33</sup> <ڤ> *ga*, and <ن> *nya*. The letter <ڤ> *v* was added for representing foreign loanwords.<sup>34</sup> This brings the size of modern Jawi inventory to 37 letters (DBP, 2006). In addition, three more characters are possible due to the adjunction of the *high hamza* <ء> above *alef* <ا>, below *alef* <ا>, and above *yeh* <ي> (MS, 2012).

As in Arabic, vowel diacritics are not normally used. However, if needed Jawi has three diacritics used to indicate the short vowels: *fatha*, *damma*, and *kasra*. A major feature of the language is the use of full reduplication of the base word (Prentice, 1990). This is represented in Jawi with the Arabic numeral <٢> (“2”) as in انجڠ-انجڠ *anjeng-anjeng* (“dogs”) as a shorthand for the equivalent longer spelling انجڠ-انجڠ for the plural form of a noun انجڠ /andʒɛŋ/ (“dog”).

#### 4. Finite-state Transformations of Perso-Arabic Script

Below we provide a brief overview of the design for Perso-Arabic script normalization framework provided by the open-source Nisaba software package.<sup>35</sup> The design was partially inspired by prior formal approaches to computational modeling of Brahmic alphasyllabaries (Datta, 1984; Sproat, 2003) and, in particular, our prior work at Brahmic script normalization (Gutkin, Johny, Doctor, Wolf-Sonkin, et al., 2022; Johny, Wolf-Sonkin, Gutkin, and Roark, 2021). These approaches exploit the inherent structure which manifests itself in all the Brahmic *abugidas* in the notion of “orthographic syllable” or *akṣara* (Bright, 1999; Fedorova, 2012). In contrast to various Brahmic scripts, the Perso-Arabic *abjad* does not offer the same rigid orthographic structure. Nevertheless, a similar in nature formal approach to script normalization, designed to address the kind of Perso-Arabic script representation ambiguities outlined in §2, can be pursued. We previously showed that scripts, such as Thaana, that borrow their features from both script families are amenable to such formal analysis (Gutkin, Johny, Doctor, Wolf-Sonkin, et al., 2022).

Our script processing pipeline consists of multiple components implemented as finite-state grammars using Pynini (Gorman, 2016; Gorman and Sproat, 2021), which is a Python framework for compiling grammars expressed as strings, regular expressions, and context-dependent rewrite rules into (weighted) finite-state transducers (FSTs).

33. The letter *fa* <ڤ> was used to represent *pa* because the sound /f/ does not exist in Malay and was pronounced as /p/.

34. The letter *va* <ڤ> is mostly used to spell English loanwords, e.g., اونيفرسيتي (“universiti”).

35. For more detailed treatment of this software please see Gutkin, Johny, Doctor, Roark, et al. (2022).

TABLE 10. Summary of script transformation operations

| Operation Type | FST             | Language-dependent | Includes        |
|----------------|-----------------|--------------------|-----------------|
| NFC            | $\mathcal{N}$   | no                 | —               |
| Common Visual  | $\mathcal{V}_c$ | no                 | $\mathcal{N}$   |
| Visual         | $\mathcal{V}$   | yes                | $\mathcal{V}_c$ |
| Reading        | $\mathcal{R}$   | yes                | $\mathcal{V}$   |

The resulting FSTs can then be efficiently combined together in a single pipeline in a variety of downstream applications (Mohri, 1996; 2009). These component FSTs are shown in Table 10 and described below.

#### Unicode Normalization

There exist language-agnostic procedures—part of the Unicode standard—that normalize text with Perso-Arabic string encodings to visually equivalent canonical normal forms. Normalization Form C (NFC) is a well-known and widely-used standard of this sort, and its application results in an equivalence class of visually identical strings that are all mapped to a single conventionalized representative of the class (Whistler, 2021). In the Nisaba library, the NFC standard is operationalized by compiling the transformations into an FST, which we denote as  $\mathcal{N}$  in Table 10. The transformations include compositions and re-orderings, along with combinations of multiple such transformations.

Composition transformations can be illustrated with the following concrete example. The *alef with madda above* letter <|> has two visually identical possible encodings: with two characters by adjoining *maddah above* to *alef* ( $\{ \text{U+0627, U+0653} \}$ ), or as the single character that already includes the maddah (U+0622). The FST  $\mathcal{N}$  transforms the two character encoding into the single character encoding, which does not change the appearance of the letter. Re-ordering transformations address multiple encodings that can arise with Arabic combining marks. As a concrete example, *shadda* (U+0651) followed by *kasra* (U+0650) yields the same rendering as *kasra* (U+0650) followed by *shadda* (U+0651). The NFC canonical form is the latter, hence the  $\mathcal{N}$  FST transforms the former encoding to the latter. The string  $\{ \text{alef (U+0627), superscript alef (U+0670), maddah above (U+0653)} \}$  is an example that transforms via  $\mathcal{N}$  with both composition and re-ordering to the visually identical form  $\{ \text{alef with madda above (U+0622), superscript alef (U+0670)} \}$ .

As noted above,  $\mathcal{N}$  is language-agnostic, meaning that its transformations (taken from the NFC standard) do not violate any language’s writing system rules.

#### Visual Normalization

We use the term *visual* normalization—initially introduced in the context of Brahmic script normalization (Johny, Wolf-Sonkin, Gutkin, and

TABLE 11. Example Urdu components included in language-specific FST  $\mathcal{V}_l$ 

| Kind of rewrite      | FST               | Letter                        | Variant (source)            | Canonical (target) |
|----------------------|-------------------|-------------------------------|-----------------------------|--------------------|
| position-independent | $\mathcal{V}_l^*$ | $\langle j \rangle$           | <i>reb + small high tab</i> | <i>rreb</i>        |
| non-final            | $\mathcal{V}_l^n$ | $\langle \mathcal{S} \rangle$ | <i>kaf</i>                  | <i>kebeb</i>       |
| word-final           | $\mathcal{V}_l^f$ | $\langle \mathcal{S} \rangle$ | <i>alef maksura</i>         | <i>farsi yeb</i>   |
| isolated-letter      | $\mathcal{V}_l^i$ | $\langle o \rangle$           | <i>beb</i>                  | <i>beb goal</i>    |

Roark, 2021)—to denote transformations that are not part of NFC but that also result in canonical forms that are visually identical to the input. This is implemented via two FSTs, one for language-agnostic transformations and one language-specific, which are combined (via FST composition) with NFC into a single language-dependent FST:  $\mathcal{V} = \mathcal{N} \circ \mathcal{V}_c \circ \mathcal{V}_l$ , where  $\circ$  denotes the composition operation (Mohri, 2009).<sup>36</sup>

The language-agnostic FST,  $\mathcal{V}_c$ , consists of the small set of normalizations not included in NFC that apply to all supported languages. As a concrete example of this class of transformations, the two-character encodings of *warw* (U+0648) followed by either *damma* (U+064F) or *small damma* (U+0619) are mapped to *u* (U+06C7). Perso-Arabic “presentation forms” from Unicode Block A, which include ligatures and contextual forms for letter variants required by the writing systems for Persian, Urdu, Sindhi and Central Asian languages,<sup>37</sup> are also normalized to visually identical canonical forms by  $\mathcal{V}_c$ , as specified by Unicode NFKC normalization (Whistler, 2021). For example, letter *beeb isolated form*  $\langle \mathcal{P} \rangle$  (U+FB52) is normalized to *beeb* (U+067B), which is visually identical. The character *ligature lam with alef isolated form*  $\langle \mathcal{L} \rangle$  (U+FEFB) is transformed to two characters: *lam*  $\langle j \rangle$  (U+0644) followed by *alef*  $\langle i \rangle$  (U+0627).

Language-specific visually-invariant transformations, included in the FST denoted as  $\mathcal{V}_l$ , include four special cases related to positions in the word: word-final, non-final (i.e., word-initial and word-medial), isolated-letter and position-independent transformations. Each of these are compiled into their own FST, as shown in Table 11, then composed into a single  $\mathcal{V}_l = \mathcal{V}_l^i \circ \mathcal{V}_l^f \circ \mathcal{V}_l^n \circ \mathcal{V}_l^*$ . Table 11 additionally presents some example transformations of each type, taken from the set of transformations required for Urdu.

### Reading Normalization

Gutkin, Johny, Doctor, Wolf-Sonkin, et al. (2022) noted the need for some additional normalization beyond those preserving visual identity for the Brahmic scripts, which they termed *reading* normalization.

36. Johny, Wolf-Sonkin, Gutkin, and Roark (2021) provides details regarding composition and other operations used by FSTs in these normalizers.

37. <https://www.unicode.org/charts/PDF/UFB50.pdf>



We also include this class of normalizations for Perso-Arabic, which we compile into the FST denoted  $\mathcal{R}$  in Table 10. Full reading normalization is the finite-state composition of visual normalization with language-specific reading normalization:  $\mathcal{R} = \mathcal{V} \circ \mathcal{R}_l$ . For example, Persian, Shahmukhi, Kashmiri, Urdu and Sorani Kurdish all map from *yeb* <ﻲ> (U+064A) to *farsi yeb* <ﻲ> (U+06CC), while Uyghur, Sindhi and Malay employ the inverse of this transformation, as dictated by their respective orthographies.

## 5. Experiments

While outlining the potential issues that may arise with text written in the Perso-Arabic script is important, it is also useful to assess how common the issues may be in real-world text. To that end, we devised some experiments that derive natural language models from collected text and validate their quality both with and without normalization. If the phenomena being normalized are rare, then the difference between the conditions will be small; and if the normalizations do not result in better text representations, then the normalized conditions may exhibit a lower quality in the validation. In this section, we present the details of our assessment, first for statistical language modeling, which provides an intrinsic validation of model quality, followed by machine translation, which provides an extrinsic validation. As was mentioned in the introduction, the code for the experiments and the corresponding results for both validation types have been released (see §1 on page 317).

### 5.1. Language Modeling Experiments

For language modeling experiments we use Wikipedia data for eight languages: five Indo-European—Kashmiri, Kurdish (Sorani), Punjabi, Sindhi, and Urdu; Malay from an Austronesian group; and Uyghur and Azerbaijani from the Turkic group. A brief overview of experimental methodology is given in §5.1.1. The dataset preprocessing details are provided in §5.1.2. The details and results of statistical language modeling experiments can be found in §5.1.3.

#### 5.1.1. Methodology

Language models are trained to predict the next token in a sequence given the previous tokens. Tokens can be variously defined as characters or words, or even as morepheme-like sub-word multi-character tokens. The intrinsic quality of the language model can be measured via the probability the language model assigns to attested exemplars, i.e.,

real text. The higher the aggregated probability of the attested text, the better the language model. See, e.g., Rosenfeld (2000) for more details on this long-standing validation paradigm. For this work, a key consideration is *comparability*—we need to ensure that models have access to the same training data and that the validation data is identical.

To ascertain whether script normalization has any significant impact on language model quality we follow a simple methodology. We adopt a  $k$ -fold cross-validation design where, for each language, we randomly shuffle the dataset and split it into the 80% training and 20% test folds, repeating the process  $k$  times, where  $k = 100$ . At each iteration we train the models (e.g., language models as in §5.1.3) and evaluate them by computing the corresponding metric (e.g., cross-entropy).

Following the above procedure, statistics are assembled using  $k$  observations for the baseline configurations that correspond to the original text and the actual testing configurations corresponding to the normalized text. Crucially, we start by generating the normalized data, recording all the sentences which contain the actual diffs in set  $D$ . During the generation of the training and test data for the baseline and testing configurations, we make sure that the sentences in  $D$  are confined to the training set. In other words, we make sure that all the actual rewrites are confined to the training data for all the  $k$  folds, which ensures that the test folds are always identical for normalized and unnormalized conditions. Given the baseline and the test metric distributions, we employ significance testing to validate the null hypothesis that the two distributions are identical; in other words, that the normalization has no significant impact on the model performance.

Three types of statistical hypothesis tests are used here. Assuming that the two groups are normally distributed an obvious choice is the two-sample (independent)  $t$ -test for comparing the means of the two populations (Zabell, 2008). Making an additional assumption that the population variances are not equal, we employ Welch’s formulation of  $t$ -test (Welch, 1947) with Satterthwaite’s degrees of freedom (Satterthwaite, 1946), referred to below as Welch-Satterthwaite (WS) test. The test provides the  $t$  statistic, the  $p$ -value and the estimated confidence interval (CI)  $[L, H]$  for the 95% confidence level at the significance level of  $\alpha = 0.05$ .

In addition, two non-parametric approaches are used here. A Mann-Whitney (MW) test (Mann and Whitney, 1947) and a more recent Brunner-Munzel (BM) test (Brunner and Munzel, 2000). Both tests provide the  $t$  statistic and the  $p$ -value. The rationale for using multiple

TABLE 12. Details of preprocessed Wikipedia datasets

| Language            | Code | $\beta$ | $N_l$     | $N_w$   | $N_l'$  | $R_l(\%)$ | $N_w'$ | $R_w(\%)$ |
|---------------------|------|---------|-----------|---------|---------|-----------|--------|-----------|
| Kashmiri            | ks   | 0.6     | 3266      | 9721    | 530     | 16.22     | 442    | 4.55      |
| Kurdish (Sorani)    | ckb  | 0.8     | 839 750   | 794 475 | 42 437  | 5.05      | 67 569 | 8.5       |
| Malay               | ms   | 0.0     | 102 311   | 200 052 | 72 645  | 71.0      | 25 854 | 12.92     |
| Punjabi (Shahmukhi) | pnb  | 0.8     | 1 075 820 | 886 399 | 145 391 | 13.51     | 16 443 | 1.86      |
| Sindhi              | sd   | 0.8     | 201 345   | 240 591 | 138 839 | 68.96     | 64 006 | 26.6      |
| South Azerbaijani   | azb  | 0.8     | 1 638 622 | 735 986 | 60 094  | 3.67      | 23 834 | 3.32      |
| Uighur              | ug   | 0.1     | 110 344   | 376 307 | 3600    | 3.26      | 19 461 | 5.17      |
| Urdu                | ur   | 0.9     | 3 595 095 | 799 610 | 6600    | 0.18      | 3632   | 0.45      |

hypothesis tests is to see whether they all agree with other providing additional weight to the null (or alternative) hypothesis.<sup>38</sup>

### 5.1.2. Corpus Preprocessing

The process of preparing the Wikipedia data is kept simple. The datasets for each language are downloaded in the MediaWiki XML format. The particular version of the dump is restricted to the pages with their current versions including the metadata.<sup>39</sup> The key difficulty lies in extracting the actual plain text in native language from the structured XML data while weeding out the metadata. We use the `mwxml` Python package developed by the Wikipedia foundation to iterate over the articles in MediaWiki XML dump.<sup>40</sup>

For each article, we use the MediaWiki Parser from Hell package to parse the current revision of article's text.<sup>41</sup> Once the parse is complete, we strip the contents of all the "unprintable" content, such as templates, using the API provided by the `mwparserfromhell` package, and split the text by newlines. A simple script detection and filtering algorithm is used to decide whether to keep the sentence or drop it from the resulting

38. All the algorithms are provided by the open-source <https://docs.scipy.org/doc/scipy/reference/stats.html> `scipy.stats` and <https://www.statsmodels.org/stable/index.html> `statsmodels` packages.

39. For example, a reasonably recent dump for Punjabi (Shahmukhi) is available at <https://dumps.wikimedia.org/pnbwiki/20211120/pnbwiki-20211120-pages-meta-current.xml.bz2>.

40. [https://www.mediawiki.org/wiki/Mediawiki\\_utilities/mwxml](https://www.mediawiki.org/wiki/Mediawiki_utilities/mwxml)

41. <https://github.com/earwig/mwparserfromhell>

data.<sup>42</sup> Any given  $l$ -character long sentence is dropped from the data if it contains less than  $\beta \cdot l$  characters in native (Perso-Arabic) script, where  $\beta \in [0, 1)$ . The filtering factor  $\beta$  is language-specific and is determined by informally examining the data.<sup>43</sup> This filtering process is crude in that it excludes any control for sentence or token length.

The preprocessing details are shown in Table 12. Each language is shown along with its Wikipedia code, the script filtering factor  $\beta$  described above, the number of resulting lines  $N_l$  and the number of unique tokens  $N_w$ . For the corresponding normalized text,  $N_l^r$  denotes the number of lines that contain diffs and  $N_w^r$  is the number of unique tokens that differ from the unnormalized version. The ratios between modified lines and token types are denoted  $R_l$  and  $R_w$ , respectively. The tokenization process is relatively crude and involves splitting on the whitespace completely disregarding other types of punctuation, such as Perso-Arabic punctuation symbols. We refer to the output of tokenization as tokens, rather than words, because the data is quite noisy even after filtering.

When normalizing text we apply the Nisaba *reading* normalization grammar (Gutkin, Johny, Doctor, Wolf-Sonkin, et al., 2022), which subsumes all the grammars providing visual invariant transformations, i.e., NFC and *visual* normalization (Johny, Wolf-Sonkin, Gutkin, and Roark, 2021), as well as transformations that change the visual appearance of the Perso-Arabic tokens. According to Table 12, the normalization effects vary across languages. For Urdu, which is the largest dataset, the percentage of modified lines and token types is below one percent. This reflects the relatively low number of transformations currently enabled in the Nisaba Urdu grammars compared to the other six languages. The highest proportion of modified lines and tokens happens in Sindhi and Malay, while for Kashmiri (the smallest datasets) and Punjabi (Shahmukhi, the second largest) the number of modifications is relatively low. A description of how these modifications affect model quality follows next.

### 5.1.3. Statistical Language Models

For building  $n$ -gram language models we use the KenLM toolkit (Heafield, 2011)<sup>44</sup> which is fast and easy to use relative to alternatives. We used modified Kneser-Ney modeling options, as recommended (Heafield, Pouzyrevsky, Clark, and Koehn, 2013b). In what follows, the terminology introduced in §5.1.1 is used. The experiments with

42. We previously implemented a similar script detection algorithm for the Wikipron project <https://github.com/CUNY-CL/wikipron>.

43. For Malay we use  $\beta = 0$ , i.e., no filtering.

44. <https://github.com/kpu/kenlm>

TABLE 13. Character-level statistics for  $k$ -fold configurations and  $m$   $n$ -gram orders ( $k = 100$ ,  $m = 8$ )

| Language | Train         |           | Test         |           |
|----------|---------------|-----------|--------------|-----------|
|          | $\mu$         | $\sigma$  | $\mu$        | $\sigma$  |
| azb      | 148 133 643.5 | 104 532.9 | 32 841 396.5 | 104 532.9 |
| ckb      | 141 088 021.3 | 118 118.6 | 32 419 200.7 | 118 118.6 |
| ks       | 252 357.7     | 3438.0    | 63 370.3     | 3438.0    |
| ms       | 16 981 259.9  | 9569.6    | 3 394 367.1  | 9569.6    |
| pnb      | 216 833 075.2 | 248 147.7 | 54 209 912.8 | 248 147.7 |
| sd       | 36 855 104.9  | 95 212.8  | 9 212 888.1  | 95 212.8  |
| ug       | 33 813 243.4  | 109 518.9 | 8 096 172.6  | 109 518.9 |
| ur       | 368 794 342.1 | 240 334.8 | 92 205 182.9 | 240 334.8 |

character and word  $n$ -gram models are described in §5.1.3 and §5.1.3, respectively.

For a single fold, the criterion for splitting into training and test sets is to use the number of lines in the corpus. As a result, the number of training and test tokens (whether these are individual characters or words) differ across the folds.

#### *Character Models*

For each language and each of the  $k = 100$  train/test folds we build  $n$ -gram character language models for orders  $n \in [3, \dots, 10]$ . The character-level statistics computed for each language over all the folds and all the orders (amounting to 800 observations per language) are shown in Table 13, where the means and standard deviations are shown for the training and test datasets. Since the corpora are split by the number of lines, the resulting variances for character datasets are quite high.

The resulting cross-entropies (in bits per character) for the models built in this way from the unnormalized text are shown in Figure 2, where each point for each  $n$ -gram order in the curve is shown along with its corresponding error band computed over  $k$  models. The plot for Kashmiri, the smallest dataset among the four languages, stands out in that the error band is clearly visible, especially for the higher orders for which the model overfits the training data. The plots for the rest of the languages show very low variance at each point in the plot for all the  $n$ -gram orders.

For each of the languages and each of the  $n$ -gram orders, we perform statistical hypothesis testing for the differences in mean cross-entropies between the character language models trained on the original (baseline, denoted  $B$ ) and the normalized (test, denoted  $T$ ) text for all the  $k$  folds ( $k = 100$ ). As mentioned in §5.1.1, for each fold, the sentences that contain (for  $T$ ) or act as source of (for  $B$ ) normalization diffs are kept in the training portion of the data. Full results for each of the languages

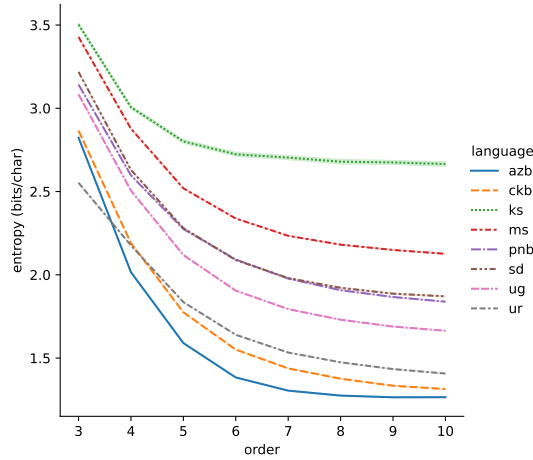


FIGURE 2. Average values of entropy (bits per character) over 100 runs vs. character  $n$ -gram orders

are presented in Appendix B, and key summarizing values are shown for all in Table 14.

The mean cross-entropy difference  $\Delta_\mu$  is computed over all the  $k$  folds as

$$\Delta_\mu = \frac{1}{k} \sum_{i=1}^k (H_i(T_i) - H_i(B_i)), \quad (1)$$

where the negative value of  $\Delta_\mu$  indicates the decrease of character entropy  $H$  of model  $i$  compared to the baseline and hence constitutes an improvement.

As can be seen from the table, the  $\Delta_\mu$  values are negative across the board, apart from the lowest  $n$ -gram orders ( $n = 3$  and  $n = 4$ ) for Sindhi. To determine whether these changes in cross-entropy are statistically significant, three types of tests (WS, MW and BM) were performed (see §5.1.1). All of the tests assess the null hypothesis that baseline and test configurations represent the same distribution. While we do not explicitly compute the correlation between the  $p$ -values for all the three tests, these tend to correlate with each other upon informal inspection. All significance test values for all languages are presented in Appendix B. Since the trends are largely the same, for ease of inspection we just show the WS  $p$ -value in Table 14, where the statistically significant degradation for Sindhi configuration corresponding to  $n = 4$  is marked in red, and discuss the few disagreements in the Appendix.

TABLE 14. Significance tests for character  $n$ -gram language models.  $\Delta_\mu$  is the mean absolute change in cross-entropy after normalization; % is the percentage change; and  $p$  is the WS  $p$ -value.

| Language            | Measure      | $n$ -gram order |        |        |        |        |        |        |        |
|---------------------|--------------|-----------------|--------|--------|--------|--------|--------|--------|--------|
|                     |              | 3               | 4      | 5      | 6      | 7      | 8      | 9      | 10     |
| South Azerbaijani   | $\Delta_\mu$ | -0.012          | -0.004 | -0.002 | -0.001 | -0.002 | -0.002 | -0.002 | -0.002 |
|                     | %            | 0.418           | 0.186  | 0.131  | 0.093  | 0.14   | 0.143  | 0.13   | 0.171  |
|                     | $p$          | 0.0             | 0.0    | 0.002  | 0.048  | 0.012  | 0.013  | 0.02   | 0.004  |
| Kurdish (Sorani)    | $\Delta_\mu$ | -0.01           | -0.003 | -0.004 | -0.003 | -0.004 | -0.006 | -0.005 | -0.005 |
|                     | %            | 0.36            | 0.16   | 0.24   | 0.2    | 0.28   | 0.44   | 0.43   | 0.39   |
|                     | $p$          | 0.0             | 0.0    | 0.0    | 0.0    | 0.0    | 0.0    | 0.0    | 0.0    |
| Kashmiri            | $\Delta_\mu$ | -0.003          | -0.021 | -0.006 | -0.006 | -0.005 | -0.016 | -0.022 | -0.028 |
|                     | %            | 0.08            | 0.71   | 0.21   | 0.22   | 0.2    | 0.63   | 0.85   | 1.07   |
|                     | $p$          | 0.647           | 0.007  | 0.544  | 0.515  | 0.564  | 0.075  | 0.014  | 0.003  |
| Malay (Jawi)        | $\Delta_\mu$ | -0.065          | -0.062 | -0.06  | -0.064 | -0.067 | -0.07  | -0.07  | -0.07  |
|                     | %            | 1.818           | 2.036  | 2.232  | 2.526  | 2.736  | 2.885  | 2.931  | 2.922  |
|                     | $p$          | 0.0             | 0.0    | 0.0    | 0.0    | 0.0    | 0.0    | 0.0    | 0.0    |
| Punjabi (Shahmukhi) | $\Delta_\mu$ | -0.011          | -0.013 | -0.01  | -0.008 | -0.007 | -0.006 | -0.007 | -0.007 |
|                     | %            | 0.32            | 0.44   | 0.39   | 0.34   | 0.33   | 0.3    | 0.36   | 0.34   |
|                     | $p$          | 0.0             | 0.0    | 0.0    | 0.0    | 0.0    | 0.0    | 0.0    | 0.0    |
| Sindhi              | $\Delta_\mu$ | 0.001           | 0.011  | -0.018 | -0.028 | -0.016 | -0.024 | -0.015 | -0.014 |
|                     | %            | -0.03           | -0.28  | 0.47   | 0.79   | 0.46   | 0.71   | 0.45   | 0.42   |
|                     | $p$          | 0.479           | 0.0    | 0.001  | 0.0    | 0.001  | 0.0    | 0.007  | 0.015  |
| Uyghur              | $\Delta_\mu$ | -0.002          | -0.001 | -0.004 | -0.004 | -0.003 | -0.004 | -0.004 | -0.005 |
|                     | %            | 0.074           | 0.051  | 0.203  | 0.219  | 0.164  | 0.225  | 0.24   | 0.295  |
|                     | $p$          | 0.0             | 0.025  | 0.0    | 0.0    | 0.0    | 0.0    | 0.0    | 0.0    |
| Urdu                | $\Delta_\mu$ | -0.003          | -0.005 | -0.004 | -0.004 | -0.004 | -0.002 | -0.002 | -0.005 |
|                     | %            | 0.11            | 0.22   | 0.2    | 0.26   | 0.24   | 0.14   | 0.13   | 0.37   |
|                     | $p$          | 0.0             | 0.0    | 0.0    | 0.0    | 0.0    | 0.02   | 0.217  | 0.001  |

TABLE 15. Word-level statistics for  $k$ -fold configurations and  $m$   $n$ -gram orders ( $k = 100$ ,  $m = 4$ )

| Language | Train        |          | Test        |          |
|----------|--------------|----------|-------------|----------|
|          | $\mu$        | $\sigma$ | $\mu$       | $\sigma$ |
| azb      | 9 704 485.4  | 7010.9   | 2 110 255.6 | 7010.9   |
| ckb      | 10 462 576.0 | 9503.5   | 2 384 669.0 | 9503.5   |
| ks       | 23 189.4     | 250.5    | 4164.6      | 250.5    |
| ms       | 1 463 801.4  | 814.4    | 290 574.6   | 814.4    |
| pnb      | 24 690 883.4 | 13 213.2 | 3 111 613.6 | 13 213.2 |
| sd       | 4 680 624.1  | 662.0    | 74 124.0    | 662.0    |
| ug       | 2 160 932.7  | 7224.1   | 515 917.3   | 7224.1   |
| ur       | 37 234 659.5 | 23 873.1 | 9 235 049.5 | 23 873.1 |

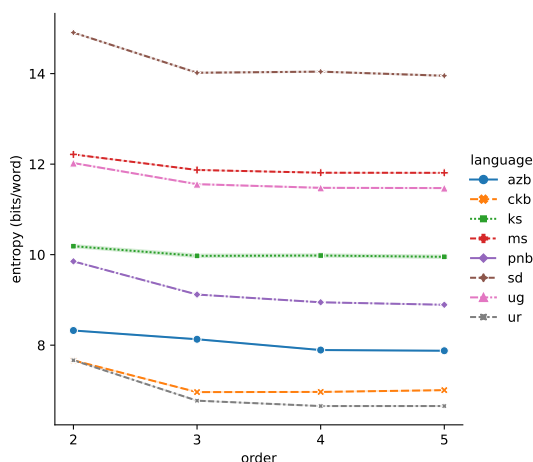


FIGURE 3. Average values of entropy (bits per word) over 100 runs vs. word  $n$ -gram orders

### Word Models

We repeated all the experiments described in §5.1.3 above for the  $n$ -gram models trained on words for the orders 2, 3, 4 and 5. The details of the training and test splits are shown in Table 15. As mentioned above, the Kashmiri dataset is very small and this sparsity is only increased when considering word-sized tokens instead of characters (compare this with Table 13). This is reinforced by computing the word cross-entropies for Kashmiri models, shown in Figure 3, where the error band shows significantly higher variance (compared to the character models in Figure 2) across each  $k$  splits for all the orders compared to other languages. Sindhi and Malay, which are the second and third smallest datasets, show a reasonably high variance as well, although it is significantly smaller than for Kashmiri. The plot for Kurdish (Sorani) indicates that the quality of the word models tends to degrade for this corpus beyond trigrams, possibly due to a relatively small size of the dataset.

Statistical significance tests were also performed for the word  $n$ -gram models constructed for  $n \in [2, 3, 4, 5]$  from the  $k = 100$  folds over original and normalized text. All values for all languages are presented in Appendix B, and key measures over all languages are shown in Table 16. Again, we just show the WS  $p$ -value in this summary table, but the values for all tests are presented and discussed in the Appendix. Kashmiri results are not significant for any of the orders, likely due to the very small size of the dataset. All other languages show statistically significant reductions in cross-entropy for all  $n$ -gram orders. Reductions are relatively small for Kurdish, Punjabi, Azerbaijani, Uyghur and Urdu, but



TABLE 16. Significance tests for word  $n$ -gram language models.  $\Delta_\mu$  is the mean absolute change in cross-entropy after normalization; % is the percentage change; and  $p$  is the WS  $p$ -value

| Language            | Measure      | $n$ -gram order |        |        |        |
|---------------------|--------------|-----------------|--------|--------|--------|
|                     |              | 2               | 3      | 4      | 5      |
| South Azerbaijani   | $\Delta_\mu$ | -0.031          | -0.027 | -0.031 | -0.028 |
|                     | %            | 0.374           | 0.332  | 0.397  | 0.358  |
|                     | $p$          | 0.0             | 0.0    | 0.0    | 0.0    |
| Kurdish (Sorani)    | $\Delta_\mu$ | -0.031          | -0.034 | -0.034 | -0.035 |
|                     | %            | 0.41            | 0.49   | 0.49   | 0.5    |
|                     | $p$          | 0.0             | 0.0    | 0.0    | 0.0    |
| Kashmiri            | $\Delta_\mu$ | 0.034           | -0.039 | -0.022 | 0.017  |
|                     | %            | -0.33           | 0.39   | 0.22   | -0.17  |
|                     | $p$          | 0.112           | 0.127  | 0.34   | 0.468  |
| Malay (Jawi)        | $\Delta_\mu$ | -0.358          | -0.394 | -0.403 | -0.411 |
|                     | %            | 2.935           | 3.319  | 3.411  | 3.479  |
|                     | $p$          | 0.0             | 0.0    | 0.0    | 0.0    |
| Punjabi (Shahmukhi) | $\Delta_\mu$ | -0.015          | -0.017 | -0.02  | -0.02  |
|                     | %            | 0.15            | 0.19   | 0.23   | 0.23   |
|                     | $p$          | 0.0             | 0.0    | 0.0    | 0.0    |
| Sindhi              | $\Delta_\mu$ | -0.159          | -0.169 | -0.177 | -0.184 |
|                     | %            | 1.06            | 1.21   | 1.26   | 1.32   |
|                     | $p$          | 0.0             | 0.0    | 0.0    | 0.0    |
| Uyghur              | $\Delta_\mu$ | -0.024          | -0.038 | -0.034 | -0.042 |
|                     | %            | 0.2             | 0.332  | 0.294  | 0.367  |
|                     | $p$          | 0.0             | 0.0    | 0.0    | 0.0    |
| Urdu                | $\Delta_\mu$ | -0.006          | -0.004 | -0.006 | -0.007 |
|                     | %            | 0.08            | 0.06   | 0.1    | 0.1    |
|                     | $p$          | 0.0             | 0.001  | 0.0    | 0.0    |

more substantial for Sindhi and particularly Malay, which achieves up to 3.5% reduction in cross-entropy.

In sum, we have shown consistently small yet significant improvements in intrinsic language model quality through the use of these normalization methods.

## 5.2. Neural Machine Translation Experiments

This section describes the application of Perso-Arabic script normalization to machine translation (MT), which is arguably one of the oldest and most popular downstream NLP tasks (Hutchins, 1986). We selected a subset of languages described in §3 and designed a simple translation experiment, where for each language we build a model that translates that source language into English in two configurations: the model trained on the original source text and the model trained on the normal-

TABLE 17. Parallel corpora details for the four languages. The dataset sizes for each language correspond to the total number of parallel sentence pairs in all datasets for a particular language.

| Languages | Train  | Development                          | Test                                     |
|-----------|--|--------------------------------------|--|
| Kurdish   | XLENT (ckb), TICO-19 (ckb),<br>Wikimedia (ckb), OPUS-100 (kur),<br>TANZIL (kur), TATOEBEA (kur)<br>256,909 sent. pairs | OPUS-100 (kur)<br><br>2,000 sent. p. | OPUS-100 (kur)<br><br>2,000 sent. p.     |
| Sindhi    | CCMATRIX, XLENT, TANZIL,<br>QED, Wikimedia<br>1,960,022 sent. pairs  | Ubuntu<br><br>6,204 sent. p.         | CCMATRIX<br>(held-out)<br>1,000 sent. p. |
| Urdu      | OPUS-100, JOSHUA, ANUVAAD<br>798,574 sent. pairs   | OPUS-100<br>2,736 sent. p.           | OPUS-100<br>2,000 sent. p.               |
| Uyghur    | XLENT, TANZIL, TATOEBEA,<br>TED, OPUS-100<br>176,179 sent. pairs   | OPUS-100<br>2,000 sent. p.           | OPUS-100<br>2,000 sent. p.               |

ized source text. We hypothesize that if the normalization is “useful,” it will result in a better model of the source language (by removing the extrinsic orthographic artifacts of Perso-Arabic resulting in systematic ambiguities) and, consequently, a better translation quality into English as measured by the objective evaluation metrics.

In what follows we introduce the parallel language corpora used for training and evaluating the models in §5.2.1, provide brief summary of the monolingual and multilingual model architectures used in §5.2.2, and discuss our results in §5.2.3. It is important to note that our aim here is not to produce a competitive MT system using current state-of-the-art (such as Wenzek et al., 2021; Xue, Barua, et al., 2022; Xue, Constant, et al., 2021), but rather to measure the effects of script normalization using reasonably advanced yet simple-to-train neural models.

### 5.2.1. *Parallel Corpora*

In our experiments we construct individual models for translating from four languages into English: Kurdish, Sindhi, Urdu and Uyghur. These parallel corpora were collected using the MTDATA tool that automates the collection and preparation of machine translation datasets (Gowda, Zhang, C. Mattmann, and May, 2021).<sup>45</sup> Each language may have multiple datasets available from several sources, such as the OPUS collection that provides various machine translation corpora for many lan-

45. <https://github.com/thammegowda/mtdata>

guages (Tiedemann, 2012).<sup>46</sup> The details of the corpora including all the datasets involved in training and testing the models for each language are shown in Table 17.

#### *Kurdish (including Sorani)*

The amount of Sorani-specific parallel Sorani-English data available online is rather small, therefore we opted to also include general Kurdish-English parallel corpora (available under kur macrolanguage ISO 639-3 code) in our training data. The resulting model is in essence a multi-dialect and multi-script model for translating from Sorani (ckb), Kurmanji (kmr), or Northern Kurdish (that uses Latin script), and possibly other Kurdish dialects into English, but we make sure that we evaluate the model using test sentence pairs that include a substantial proportion of source sentences in Perso-Arabic script.

The training set includes the Sorani (ckb) part of XLENT (El-Kishky et al., 2021), TICO-19 (Anastasopoulos et al., 2020),<sup>47</sup> and Wikimedia (Tiedemann, 2012),<sup>48</sup> as well as the general Kurdish (kur) training portions of OPUS-100 (Zhang, Williams, Titov, and Sennrich, 2020),<sup>49</sup> the TANZIL corpus of religious texts,<sup>50</sup> and the TATOEBa dataset.<sup>51</sup> The development set consists of the development portion of OPUS-100 for general Kurdish. The test set is a general Kurdish test set of OPUS-100 dataset. In this set there are 329 source (Kurdish) sentences out of 2,000 which are in Perso-Arabic script.

#### *Sindhi*

The Sindhi training data includes the Sindhi-English parallel data from the following datasets: XLENT, Wikimedia, TANZIL, CCMATRIX (Fan et al., 2021; Schwenk et al., 2021),<sup>52</sup> and QED (Abdelali, Guzman, Sajjad, and Vogel, 2014).<sup>53</sup> For the development set we selected the Sindhi-English localization strings from the Ubuntu project.<sup>54</sup> The test set consists of 1,000 sentence pairs withheld from the OPUS-100 training set.

---

46. <https://opus.nlpl.eu/>

47. <https://opus.nlpl.eu/tico-19.php>

48. <https://opus.nlpl.eu/wikimedia.php>

49. <https://opus.nlpl.eu/opus-100.php>

50. <https://opus.nlpl.eu/Tanzil.php> (<https://tanzil.net/>)

51. <https://opus.nlpl.eu/Tatoeba.php> (<https://tatoeba.org/en/>)

52. <https://opus.nlpl.eu/CCMatrix.php>

53. <https://opus.nlpl.eu/QED.php>

54. <https://opus.nlpl.eu/Ubuntu.php>

### *Urdu*

In the training data we include the Urdu-English sentence pairs from the ANUVAAD corpus (Anuvaad, 2022), the parallel South Asian corpora from the JOSHUA statistical machine translation (SMT) toolkit (Post, Callison-Burch, and Osborne, 2012),<sup>55</sup> and the training set from OPUS-100. The development and test sets consist of the respective development and test Urdu-English partitions of the OPUS-100 dataset.

### *Uyghur*

The training data includes the Uyghur-English pairs from the following corpora: XLENT, TANZIL, TATOEBEA, TED (Reimers and Gurevych, 2020),<sup>56</sup> and the training partition of OPUS-100. The development and test sets consist of the respective development and test Uyghur-English partitions of the OPUS-100 dataset.

### *Multilingual Configuration*

In addition to constructing individual monolingual translation models we also experiment with a single multilingual model that provides many-to-English translation. Rather than using all the available data we constructed a corpus that is balanced in terms of per-language parallel sentence pairs: all of the training data is selected for Uyghur, which is our smallest dataset (see Table 17), and for the rest of the languages we selected the first 200,000 sentence pairs from the respective training sets. The resulting multilingual training set thus constructed consists of 776,179 sentence pairs. The test set for multilingual configuration consists of 7,000 sentence pairs that correspond to the whole test sets for the respective languages.

### 5.2.2. *Models*

Modern neural machine translation (NMT) models are an instance of neural sequence-to-sequence models, which have achieved impressive results in recent years (Stahlberg, 2020). In our experiments we use a variant of recurrent neural network (RNN) encoder-decoder bipartite architecture equipped with attention mechanism (Bahdanau, Cho, and Bengio, 2015; Mnih, Heess, Graves, and Kavukcuoglu, 2014), where instead of RNN units, long short-term memory (LSTM) cells are used, which allows the network to learn the long sequences more efficiently (Hochreiter and Schmidhuber, 1997). The particular attention mechanism we use in the decoder is described in Luong, Pham, and Manning (2015).

---

55. <https://github.com/joshua-decoder/indian-parallel-corpora>

56. <https://opus.nlpl.eu/TED2020.php>

We use two different model configurations. For languages with larger amount of data (Sindhi and Urdu, as shown in Table 17), the encoder component is bidirectional (Schuster and Paliwal, 1997), consisting of four stacked layers of 256 LSTM units each, while the decoder memory consists of 512 units. The configuration used for languages with smaller amounts of data, Kurdish and Uyghur, is mostly identical, but the encoder has two unidirectional LSTM layers. Both models correspond to vanilla configurations (NMTMediumV1 and NMTSmallV1) provided by the OPENNMT-TF library (Klein et al., 2017) implemented in the TensorFlow framework (Abadi et al., 2016). Our models are word-based, with the 50,000 most frequent words used for source and target embedding vocabularies. The tokenization is performed in *aggressive* mode provided by the default OPENNMT tokenizer. The parallel sentence pairs where either the source or the target sentence is longer than 100 words are dropped from the training. Overall, the larger models NMTMediumV1 have approximately 92M model parameters, while the smaller models NMTSmallV1 have approximately 62M parameters. We used default hyperparameters provided by the OPENNMT configurations, apart from the training batch size which we set to 64 examples.

For the multilingual experiment, the size of the balanced dataset described in §5.2.1 roughly corresponds to the size of our Urdu corpus. Hence, similar to Urdu and Sindhi, we have chosen the NMTMediumV1 configuration for our multilingual many-to-English model with the same hyperparameters as for the monolingual configurations.

### 5.2.3. Results and Discussion

For each language two native language-to-English models were trained from unnormalized and normalized text for that language, respectively.<sup>57</sup> The details of the language-specific text partitions are provided in §5.2.1. Perso-Arabic script normalization was applied to the native language side of training, development and testing portions of the data, with the English side kept unchanged. Each model was trained for 8 epochs and at the end of each epoch model’s performance was evaluated on the test set using the three MT metrics, each using default parameters such as casing and smoothing, provided by the SACRE-BLEU toolkit (Post, 2018):<sup>58</sup> the BiLingual Evaluation Understudy, or BLEU score (Papineni, Roukos, Ward, and Zhu, 2002), the Character  $n$ -gram F-score, or CHRF2 (Popović, 2016), and Translation Edit Rate,

57. While we do not provide the statistics for the normalized NMT data in terms of number of training set lines, tokens and types changed by the normalization, we hypothesize that these ratios would be similar to the ones computed for statistical language modeling experiments using Wikipedia data presented in Table 12 in §5.1.

58. <https://github.com/mjpost/sacrebleu>

TABLE 18. Relative difference (%) between the performance of normalized and unnormalized models

| (a) Kurdish |               |                |              | (b) Sindhi |               |                |              |
|-------------|---------------|----------------|--------------|------------|---------------|----------------|--------------|
| Epochs      | $\Delta$ BLEU | $\Delta$ CHRF2 | $\Delta$ TER | Epochs     | $\Delta$ BLEU | $\Delta$ CHRF2 | $\Delta$ TER |
| 1           | -3.216        | 30.460         | 0.665        | 1          | 11.239        | 2.264          | 4.059        |
| 2           | -10.719       | -5.340         | -0.883       | 2          | 6.358         | 2.351          | -3.655       |
| 3           | 2.249         | 8.010          | -3.095       | 3          | 3.536         | 0.028          | -2.644       |
| 4           | 30.132        | 14.786         | -2.996       | 4          | 7.950         | 3.158          | -2.161       |
| 5           | 28.440        | 17.847         | 1.985        | 5          | 2.024         | 1.542          | 4.414        |
| 6           | 20.165        | 16.023         | -2.022       | 6          | 1.075         | 0.652          | -0.503       |
| 7           | 17.866        | 12.143         | -6.269       | 7          | 3.384         | 3.173          | 1.117        |
| 8           | 31.357        | 18.202         | -4.183       | 8          | 1.254         | -0.298         | -2.795       |
| $\mu$       | 14.534        | 14.015         | -2.099       | $\mu$      | 4.603         | 1.609          | -0.271       |

| (c) Urdu |               |                |              | (d) Uyghur |               |                |              |
|----------|---------------|----------------|--------------|------------|---------------|----------------|--------------|
| Epochs   | $\Delta$ BLEU | $\Delta$ CHRF2 | $\Delta$ TER | Epochs     | $\Delta$ BLEU | $\Delta$ CHRF2 | $\Delta$ TER |
| 1        | -2.063        | -2.960         | -2.657       | 1          | -5.998        | -1.583         | 0.130        |
| 2        | 10.752        | 3.428          | -1.981       | 2          | 5.920         | 1.378          | 2.780        |
| 3        | 6.686         | 1.269          | -2.686       | 3          | 2.661         | 0.960          | 0.128        |
| 4        | 8.741         | 4.109          | -0.764       | 4          | 6.525         | 7.454          | -5.967       |
| 5        | 4.781         | 1.423          | 0.121        | 5          | 1.874         | 0.046          | 1.550        |
| 6        | 3.657         | 0.610          | -1.366       | 6          | -2.120        | -3.592         | -2.101       |
| 7        | 4.634         | 1.606          | -2.558       | 7          | -0.345        | 1.451          | 0.748        |
| 8        | 3.367         | 1.726          | -0.441       | 8          | 0.600         | 0.348          | 1.224        |
| $\mu$    | 5.069         | 1.401          | -1.541       | $\mu$      | 1.140         | 0.808          | -0.188       |

or TER (Snover et al., 2006). Higher BLEU and CHRF2 scores indicate that the hypotheses better match the reference translations, whereas for TER lower scores indicate a better match.

The relative differences (in %) computed between the scores for the models build on normalized and unnormalized text for each language are shown in Table 18, with the positive values of  $\Delta$  BLEU and  $\Delta$  CHRF2, and the negative values of  $\Delta$  TER signifying relative improvement in performance of the normalized model over the unnormalized one at each training epoch. The last two highlighted rows in each table correspond to relative performance differences at the last epoch and the mean per-epoch difference  $\mu$ . The improvements are highlighted in green and the degradation in red.<sup>59</sup>

As can be seen from Table 18, the biggest gains over the baseline are obtained for the normalized Kurdish model over all the three MT metrics for both the last training epoch as well as the average per-epoch relative difference. The normalized Urdu model also displays improvements across the board. For Sindhi, there is a relative degradation of 0.298%

59. The absolute raw scores are also provided in Table 19.

TABLE 19. Paired significance tests for the monolingual models obtained after the final training epoch: Paired Bootstrap Resampling (PBS) and Paired Approximate Randomization (PAR)

| Systems |               | BLEU ( $\mu \pm 95\%$ CI)   | PBS                         |                             |        | PAR    |        |  |
|---------|---------------|-----------------------------|-----------------------------|-----------------------------|--------|--------|--------|--|
|         |               |                             | CHRF2 ( $\mu \pm 95\%$ CI)  | TER ( $\mu \pm 95\%$ CI)    | BLEU   | CHRF2  | TER    |  |
| Kurdish | $\mathcal{B}$ | 10.740 (10.731 $\pm$ 1.082) | 27.187 (27.193 $\pm$ 1.068) | 83.047 (83.070 $\pm$ 1.442) | 10.740 | 27.187 | 83.047 |  |
|         | $\mathcal{N}$ | 15.646 (15.617 $\pm$ 1.269) | 33.237 (33.220 $\pm$ 1.210) | 79.712 (79.742 $\pm$ 1.769) | 15.646 | 33.237 | 79.712 |  |
|         | $f$           | 0.001                       | 0.001                       | 0.001                       | 0.000  | 0.000  | 0.000  |  |
| Sindhi  | $\mathcal{B}$ | 15.362 (15.367 $\pm$ 0.675) | 39.223 (39.228 $\pm$ 0.707) | 74.960 (74.967 $\pm$ 1.034) | 15.362 | 39.223 | 74.960 |  |
|         | $\mathcal{N}$ | 15.557 (15.572 $\pm$ 0.717) | 39.106 (39.116 $\pm$ 0.719) | 72.921 (72.923 $\pm$ 0.991) | 15.557 | 39.106 | 72.921 |  |
|         | $f$           | 0.148                       | 0.195                       | 0.001                       | 0.392  | 0.575  | 0.000  |  |
| Urdu    | $\mathcal{B}$ | 13.387 (13.393 $\pm$ 0.702) | 33.559 (33.562 $\pm$ 0.771) | 77.403 (77.402 $\pm$ 1.002) | 13.387 | 33.559 | 77.403 |  |
|         | $\mathcal{N}$ | 13.854 (13.854 $\pm$ 0.709) | 34.148 (34.151 $\pm$ 0.751) | 77.064 (77.061 $\pm$ 1.066) | 13.854 | 34.148 | 77.064 |  |
|         | $f$           | 0.043                       | 0.005                       | 0.167                       | 0.091  | 0.010  | 0.442  |  |
| Uyghur  | $\mathcal{B}$ | 9.982 (9.938 $\pm$ 0.588)   | 29.118 (29.117 $\pm$ 0.636) | 87.236 (87.207 $\pm$ 1.678) | 9.982  | 29.118 | 87.236 |  |
|         | $\mathcal{N}$ | 10.043 (10.017 $\pm$ 0.581) | 29.220 (29.225 $\pm$ 0.628) | 88.317 (88.352 $\pm$ 1.413) | 10.043 | 29.220 | 88.317 |  |
|         | $f$           | 0.315                       | 0.225                       | 0.105                       | 0.825  | 0.668  | 0.277  |  |

in CHRF2 over the unnormalized baseline at the last training epoch, while for Uyghur there is a larger final-epoch degradation of 1.224% in TER. Apart from these two cases however, overall the mean per-epoch and final-epoch relative differences indicate potential improvements, although in the case of Uyghur these are small.

In order to ascertain whether the above relative differences are statistically significant we performed paired significance testing of the final-epoch systems using two algorithms provided by SACREBLEU: the Paired Bootstrap Resampling (Koehn, 2004) and the Paired Approximate Randomization (Riezler and Maxwell, 2005), denoted PBS and PAR, respectively. For PBS, the default parameter of 1,000 resampling trials was used. For PAR, the default value of 10,000 trials was used for randomization test. The results of both tests are shown in Table 19. For each language two systems are tested: the unnormalized baseline ( $\mathcal{B}$ ), and the model built from the normalized text ( $\mathcal{N}$ ). The systems are pairwise compared for sentences from the test set using the three MT metrics described above. The null hypotheses for both tests postulate that both  $\mathcal{B}$  and  $\mathcal{N}$  translations are generated by the same underlying process. For a given model  $\mathcal{N}$  and the baseline  $\mathcal{B}$ , the  $p$ -value is roughly the probability of the absolute score difference ( $\Delta$ ) or higher occurring due to chance, under the assumption that the null hypothesis is correct. Assuming a significance threshold of 0.05, the null hypothesis can be rejected for  $p$ -values  $< 0.05$ , which implies that both systems are different. For PBS, the actual system score, the bootstrap estimated true mean ( $\mu$ ), and the 95% confidence interval (CI) are shown for each metric. For PAR, no true mean or confidence intervals are shown because the algorithm does not perform the resampling.

In Table 19 the statistically significant improvements are highlighted in green, while the cases where the systems appear to be equivalent are

TABLE 20. Relative difference (%) between the performance of two (normalized and unnormalized) many-to-English models

| Epochs | $\Delta$ BLEU | $\Delta$ CHRF2 | $\Delta$ TER |
|--------|---------------|----------------|--------------|
| 1      | 31.456        | 12.001         | 5.988        |
| 2      | 35.790        | 18.781         | -4.837       |
| 3      | 28.537        | 14.952         | -2.812       |
| 4      | 25.456        | 13.398         | -5.170       |
| 5      | 22.387        | 9.781          | -4.486       |
| 6      | 14.101        | 7.564          | -6.108       |
| 7      | 15.732        | 8.336          | -4.087       |
| 8      | 10.666        | 5.702          | -6.660       |
| $\mu$  | 23.016        | 11.314         | -3.521       |

highlighted in light blue. We note that both small degradations in translation quality of Sindhi and Uyghur for the individual metrics observed in Table 18 turn out to be not statistically significant, as evidenced by the corresponding  $p$ -values. For Kurdish, the improvements are statistically significant across the board, while for Sindhi and Urdu the improvements are significant according to at least one MT metric and at least one significance test: both PBS and PAR agree on improvements in TER on Sindhi and in CHRF2 on Urdu (where PBS also indicates significant improvement in BLEU). Interestingly, both tests indicate that normalization has no effect on Uyghur translation quality. We hypothesize that this may be due to several conflating factors. First, since this is the smallest dataset of all the languages in this experiment (see Table 17), there may not be enough data for training the model reliably. Furthermore, potential misalignment between Uyghur and English sentences in the training data may be adversely affecting the quality of the resulting models.

#### *Many-to-English Experiment*

The goal of this experiment is to verify the hypothesis that the relative performance of Perso-Arabic script-normalized individual NMT systems, especially Uyghur, is improved by pooling the data from other available languages. To this end we trained a single many-to-English model described in §5.2.1 and §5.2.2. Similar to individual language-to-English experiments, we compare the performance of the NMT model built from normalized text against the baseline model constructed from unnormalized data.

Before proceeding two important points need to be noted. First, because our Perso-Arabic script normalization grammars are language-specific, the normalized version of the multilingual corpus described in §5.2.1 is constructed from the normalized corpora for the respective individual languages. Second, since our balanced multilingual corpus



TABLE 21. Paired significance tests for the multilingual model obtained after the final training epoch: Paired Bootstrap Resampling (PBS) and Paired Approximate Randomization (PAR)

| Systems |                 | PBS                         |                             |                             | PAR    |        |        |
|---------|-----------------|-----------------------------|-----------------------------|-----------------------------|--------|--------|--------|
|         |                 | BLEU ( $\mu \pm 95\%$ CI)   | CHRF2 ( $\mu \pm 95\%$ CI)  | TER ( $\mu \pm 95\%$ CI)    | BLEU   | CHRF2  | TER    |
| Kurdish | $\mathcal{B}^m$ | 12.968 (12.891 $\pm$ 1.471) | 29.403 (29.412 $\pm$ 1.144) | 83.296 (83.300 $\pm$ 2.881) | 12.968 | 29.402 | 83.296 |
|         | $\mathcal{N}^m$ | 18.496 (18.505 $\pm$ 1.511) | 34.773 (34.781 $\pm$ 1.334) | 73.373 (73.371 $\pm$ 1.820) | 18.496 | 34.773 | 73.373 |
|         | $p$             | 0.001                       | 0.001                       | 0.001                       | 0.000  | 0.000  | 0.000  |
| Sindhi  | $\mathcal{B}^m$ | 14.602 (14.586 $\pm$ 0.690) | 37.889 (37.901 $\pm$ 0.649) | 76.918 (76.920 $\pm$ 1.105) | 14.602 | 37.889 | 76.918 |
|         | $\mathcal{N}^m$ | 15.715 (15.727 $\pm$ 0.742) | 39.410 (39.425 $\pm$ 0.675) | 72.889 (72.895 $\pm$ 1.011) | 15.715 | 39.410 | 72.890 |
|         | $p$             | 0.001                       | 0.001                       | 0.001                       | 0.000  | 0.000  | 0.000  |
| Urdu    | $\mathcal{B}^m$ | 11.226 (11.211 $\pm$ 0.621) | 30.516 (30.526 $\pm$ 0.677) | 84.246 (84.236 $\pm$ 1.651) | 11.226 | 30.516 | 84.246 |
|         | $\mathcal{N}^m$ | 12.255 (12.247 $\pm$ 0.658) | 32.057 (32.057 $\pm$ 0.721) | 79.169 (79.185 $\pm$ 1.065) | 12.255 | 32.057 | 79.169 |
|         | $p$             | 0.001                       | 0.001                       | 0.001                       | 0.000  | 0.000  | 0.000  |
| Uyghur  | $\mathcal{B}^m$ | 15.346 (15.361 $\pm$ 0.820) | 35.412 (35.419 $\pm$ 0.788) | 75.741 (75.731 $\pm$ 1.206) | 15.346 | 35.412 | 75.741 |
|         | $\mathcal{N}^m$ | 17.031 (17.050 $\pm$ 0.881) | 37.377 (37.388 $\pm$ 0.877) | 71.677 (71.656 $\pm$ 1.182) | 17.031 | 37.377 | 71.677 |
|         | $p$             | 0.001                       | 0.001                       | 0.001                       | 0.000  | 0.000  | 0.000  |

has roughly the same number of sentence pairs for each language, the amount of within-language data available for the “bigger” languages (Urdu and Sindhi) is significantly smaller in this experiment. Thus, when one compares NMT scores for these languages between a many-to-English system on one hand, and a particular monolingual language-to-English system on the other, the many-to-English scores may be worse. This is not an issue because, as mentioned above, the goal of this experiment is to investigate the relative improvements over the unnormalized baseline, rather than constructing an NMT system with the best possible absolute score.

The relative differences (in %) computed between the scores for the many-to-English models build from normalized and unnormalized text are shown in Table 20. It is worth noting that unlike the monolingual scores shown in Table 18 these scores are computed using the combined test data consisting of 7,000 sentence pairs from all the individual languages described in §5.2.1. As can be seen from Table 20, the normalized many-to-English model shows consistent improvements in all the metrics over the unnormalized baseline for all the epochs with the exception of 5.988% degradation in TER for the initial epoch.

We also performed PBS and PAR paired statistical significance tests for the many-to-English configuration comparing the performance of the multilingual normalized model (denoted  $\mathcal{N}^m$ ) against its unnormalized counterpart ( $\mathcal{B}^m$ ) on the test data for individual languages. The results of both tests are shown in Table 21 for each language, with the statistically significant improvements in individual metrics marked in green. Compared to significance tests for the monolingual systems in Table 19, the multilingual tests show more robust improvements across all languages and metrics. In particular, with respect to Uyghur these results confirm our hypothesis above that the original dataset is

too small to reliably measure the effects of script normalization. This is rectified by using data from other languages. Additional supporting evidence comes from comparing between the absolute values of all the metrics for normalized and unnormalized monolingual and multilingual models for Uyghur and Kurdish shown in Table 19 and Table 21: multilingual configurations have higher absolute scores for these languages.

## 6. Conclusions

This paper provided a brief overview of various adaptations of the Perso-Arabic script for eight languages from diverse language families and the issues that result from representing these adaptations digitally. The particular emphasis of this study was on the visual ambiguities between Perso-Arabic characters represented in Unicode. We argued that the computational methods for visual disambiguation need to go beyond the standard language-agnostic techniques provided by the Unicode standard and take into account the specifics of a local writing system as well as multiple confounding factors that affect the patterns of its use. We presented two types of writing system-specific normalization methods. Similar to the standard Unicode normalization techniques, *visual* normalization preserves the visual invariance of the characters, while providing significantly broader coverage of normalization cases peculiar to the orthography in question. The second type is *reading* normalization, which provides character transformations that violate visual invariance (e.g., by modifying the number of *ijām* dots on the base shape of a character), yet are required to make the input conform to the local orthography. The distinction between the two types hinges on the visual invariance criterion, which is helpful in deciding when and if to apply either type of technique.<sup>60</sup> Perso-Arabic script normalization techniques are crucial for cybersecurity, but the focus of this paper is on their application to natural language processing. We performed experiments in statistical language modeling and neural machine translation that demonstrated the positive impact of script normalization on the performance of the resulting models.

This study describes work that is still in early stages. While there is a wealth of literature on the eight languages described in this paper and some additional languages currently covered by our methods, the majority of Perso-Arabic writing systems are used for lower-resource languages that are either scarcely documented or have very little online data, which is needed to provide evidence for required normalization.

---

60. In some applications it may be necessary to preserve the visual fidelity of the input, hence the application of reading normalization may not be desirable.

Furthermore, significant research towards a formal description of Perso-Arabic script typology is still required.

## Acknowledgements

The authors would like to thank Christo Kirov for many useful comments on an earlier draft of this paper, as well as Lawrence Wolf-Sonkin, Işın Demirşahin and Anna Katanova for informative feedback and assistance with the various stages of this project. We also thank Aso Mahmudi for his invaluable feedback on the modern orthography of Central Kurdish.

## A. Letter Inventories for Individual Languages

The letter repertoire for the eight languages investigated in this paper—South Azerbaijani (azb), Sorani Kurdish (ckb), Kashmiri (ks), Malay (ms), Western Punjabi (pnb), Sindhi (sd), Uyghur (ug) and Urdu (ur)—is shown in Table 22 below. Overall we identified 118 characters including letters and various diacritics. The table shows, for each character its corresponding Unicode code point, Unicode name and the languages that use it (indicated by the checkmark).

TABLE 22. Letter inventories for individual languages

| Char. | Codepoint | Character Name        | Language Tags |     |    |    |     |    |    |    |
|-------|-----------|-----------------------|---------------|-----|----|----|-----|----|----|----|
|       |           |                       | azb           | ckb | ks | ms | pnb | sd | ug | ur |
| <ي>   | U+0620    | Kashmiri Yeh          |               |     | ✓  |    |     |    |    |    |
| <ء>   | U+0621    | Hamza                 |               |     |    | ✓  | ✓   | ✓  |    | ✓  |
| <إ>   | U+0622    | Alef with Madda Above | ✓             |     | ✓  | ✓  | ✓   | ✓  |    | ✓  |
| <أ>   | U+0623    | Alef with Hamza Above |               |     | ✓  | ✓  | ✓   | ✓  |    | ✓  |
| <ؤ>   | U+0624    | Waw with Hamza Above  | ✓             |     | ✓  | ✓  | ✓   | ✓  |    | ✓  |
| <آ>   | U+0625    | Alef with Hamza Below |               |     | ✓  | ✓  | ✓   |    |    | ✓  |
| <آ>   | U+0626    | Yeh with Hamza Above  | ✓             | ✓   |    |    | ✓   | ✓  | ✓  | ✓  |
| <ا>   | U+0627    | Alef                  | ✓             | ✓   | ✓  | ✓  | ✓   | ✓  | ✓  | ✓  |
| <ب>   | U+0628    | Beh                   | ✓             | ✓   | ✓  | ✓  | ✓   | ✓  | ✓  | ✓  |
| <ة>   | U+0629    | Teh Marbuta           |               |     |    | ✓  | ✓   |    | ✓  | ✓  |
| <ت>   | U+062A    | Teh                   | ✓             | ✓   | ✓  | ✓  | ✓   | ✓  | ✓  | ✓  |
| <ٹ>   | U+062B    | Theh                  | ✓             |     | ✓  | ✓  | ✓   | ✓  |    | ✓  |
| <ج>   | U+062C    | Jeem                  | ✓             | ✓   | ✓  | ✓  | ✓   | ✓  | ✓  | ✓  |
| <ح>   | U+062D    | Hah                   | ✓             | ✓   | ✓  | ✓  | ✓   | ✓  | ✓  | ✓  |
| <خ>   | U+062E    | Khah                  | ✓             | ✓   | ✓  | ✓  | ✓   | ✓  | ✓  | ✓  |

*Continued on next page*

TABLE 22—Continued from previous page

| Char. | Codepoint | Character Name                | Language Tags |     |    |    |    |    |    |    |
|-------|-----------|-------------------------------|---------------|-----|----|----|----|----|----|----|
|       |           |                               | az            | ckb | ks | ms | pa | sd | ug | ur |
| <د>   | U+062F    | Dal                           | ✓             | ✓   | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <ذ>   | U+0630    | Thal                          | ✓             |     | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <ر>   | U+0631    | Reh                           | ✓             | ✓   | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <ز>   | U+0632    | Zain                          | ✓             | ✓   | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <س>   | U+0633    | Seen                          | ✓             | ✓   | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <ش>   | U+0634    | Sheen                         | ✓             | ✓   | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <ص>   | U+0635    | Sad                           | ✓             |     | ✓  | ✓  | ✓  | ✓  |    | ✓  |
| <ض>   | U+0636    | Dad                           | ✓             |     | ✓  | ✓  | ✓  | ✓  |    | ✓  |
| <ط>   | U+0637    | Tah                           | ✓             |     | ✓  | ✓  | ✓  | ✓  |    | ✓  |
| <ظ>   | U+0638    | Zah                           | ✓             |     | ✓  | ✓  | ✓  | ✓  |    | ✓  |
| <ع>   | U+0639    | Ain                           |               | ✓   | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <غ>   | U+063A    | Ghain                         | ✓             | ✓   | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <ی>   | U+063D    | Farsi Yeh with<br>Inverted V  | ✓             |     |    |    |    |    |    |    |
| <ف>   | U+0641    | Feh                           | ✓             | ✓   | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <ق>   | U+0642    | Qaf                           | ✓             | ✓   | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <ك>   | U+0643    | Kaf                           |               |     |    |    |    |    | ✓  | ✓  |
| <ل>   | U+0644    | Lam                           | ✓             | ✓   | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <م>   | U+0645    | Meem                          | ✓             | ✓   | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <ن>   | U+0646    | Noon                          | ✓             | ✓   | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <ه>   | U+0647    | Heh                           | ✓             | ✓   |    | ✓  | ✓  |    |    | ✓  |
| <و>   | U+0648    | Waw                           | ✓             | ✓   | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <ى>   | U+0649    | Alef Maksura                  |               | ✓   |    | ✓  | ✓  |    | ✓  | ✓  |
| <ي>   | U+064A    | Yeh                           |               |     |    | ✓  |    | ✓  | ✓  |    |
| <َ>   | U+064B    | Fathatan                      |               |     |    |    | ✓  |    |    | ✓  |
| <ِ>   | U+064C    | Dammatan                      |               |     |    |    | ✓  |    |    | ✓  |
| <ُ>   | U+064D    | Kasratan                      |               |     |    |    | ✓  |    |    | ✓  |
| <ا>   | U+064E    | Fatha                         | ✓             |     | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <آ>   | U+064F    | Damma                         | ✓             |     | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <إ>   | U+0650    | Kasra                         | ✓             |     | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <ع>   | U+0651    | Shadda                        | ✓             |     | ✓  | ✓  | ✓  | ✓  |    | ✓  |
| <ـ>   | U+0652    | Sukun                         | ✓             |     | ✓  | ✓  | ✓  |    |    | ✓  |
| <ْ>   | U+0653    | Maddah Above                  | ✓             |     | ✓  | ✓  | ✓  | ✓  |    | ✓  |
| <َٕ>  | U+0654    | Hamza Above                   |               |     | ✓  |    | ✓  | ✓  | ✓  | ✓  |
| <ِٕ>  | U+0655    | Hamza Below                   |               |     | ✓  |    | ✓  | ✓  |    | ✓  |
| <ُٕ>  | U+0656    | Subscript Alef                |               |     | ✓  |    | ✓  |    |    |    |
| <إ>  | U+0657    | Inverted Damma                |               |     | ✓  |    | ✓  |    |    | ✓  |
| <ٓ>   | U+065A    | Vowel Sign Small V<br>Above   |               |     | ✓  |    |    |    |    |    |
| <ٔ>   | U+065F    | Wavy Hamza Below              |               |     | ✓  |    |    |    |    |    |
| <ٖ>   | U+0670    | Superscript Alef              |               |     |    |    | ✓  |    |    | ✓  |
| <اٲ>  | U+0671    | Alef Wasla                    | ✓             |     |    |    |    |    |    |    |
| <اٱ>  | U+0672    | Alef with Wavy<br>Hamza Above |               |     | ✓  |    |    |    |    |    |
| <اٰ>  | U+0673    | Alef with Wavy<br>Hamza Below |               |     | ✓  |    |    |    |    |    |
| <ٲ>   | U+0679    | Tteh                          |               |     | ✓  |    | ✓  |    |    | ✓  |
| <ٱ>   | U+067A    | Tteheh                        |               |     |    |    |    | ✓  |    |    |

Continued on next page

TABLE 22—Continued from previous page

| Char. | Codepoint | Character Name                      | Language Tags |     |    |    |    |    |    |    |
|-------|-----------|-------------------------------------|---------------|-----|----|----|----|----|----|----|
|       |           |                                     | az            | ckb | ks | ms | pa | sd | ug | ur |
| <ٻ>   | U+067B    | Beeh                                |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+067D    | Teh with Three Dots Above Downwards |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+067E    | Peh                                 | ✓             | ✓   | ✓  |    | ✓  | ✓  | ✓  | ✓  |
| <ٲ>   | U+067F    | Teheh                               |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+0680    | Beheh                               |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+0683    | Nyeh                                |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+0684    | Dyeh                                |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+0685    | Hah with Three Dots Above           |               |     |    |    |    |    | ✓  |    |
| <ٲ>   | U+0686    | Tcheh                               | ✓             | ✓   | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  |
| <ٲ>   | U+0687    | Tcheheh                             |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+0688    | Ddal                                |               |     | ✓  |    | ✓  |    |    | ✓  |
| <ٲ>   | U+068A    | Dal with Dot Below                  |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+068C    | Dahal                               |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+068D    | Ddahal                              |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+068F    | Dal with Three Dots Above Downwards |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+0691    | Rreh                                |               |     | ✓  |    | ✓  |    |    | ✓  |
| <ٲ>   | U+0695    | Reh with Small V Below              |               | ✓   |    |    |    |    |    |    |
| <ٲ>   | U+0698    | Jeh                                 | ✓             | ✓   | ✓  |    | ✓  |    | ✓  | ✓  |
| <ٲ>   | U+0699    | Reh with Four Dots Above            |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+06A0    | Ain with Three Dots Above           |               |     |    | ✓  |    |    |    |    |
| <ٲ>   | U+06A4    | Veh                                 |               | ✓   |    | ✓  |    |    |    |    |
| <ٲ>   | U+06A6    | Peheh                               |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+06A9    | Keheh                               | ✓             | ✓   | ✓  | ✓  | ✓  | ✓  |    | ✓  |
| <ٲ>   | U+06AA    | Swash Kaf                           |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+06AD    | Ng                                  |               |     |    |    |    |    | ✓  |    |
| <ٲ>   | U+06AF    | Gaf                                 | ✓             | ✓   | ✓  |    | ✓  | ✓  | ✓  | ✓  |
| <ٲ>   | U+06B1    | Ngoeh                               |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+06B3    | Gueh                                |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+06B4    | Gaf with Three Dots Above           | ✓             |     |    |    |    |    |    |    |
| <ٲ>   | U+06B5    | Lam with Small V                    |               | ✓   |    |    |    |    |    |    |
| <ٲ>   | U+06BA    | Noon Ghunna                         |               |     | ✓  |    | ✓  |    |    | ✓  |
| <ٲ>   | U+06BB    | Rnoon                               |               |     |    |    |    | ✓  |    |    |
| <ٲ>   | U+06BD    | Noon with Three Dots Above          |               |     |    | ✓  |    |    |    |    |
| <ٲ>   | U+06BE    | Heh Doachashmee                     | ✓             | ✓   |    |    | ✓  | ✓  | ✓  | ✓  |
| <ٲ>   | U+06C0    | Heh with Yeh Above                  | ✓             |     |    |    |    |    |    |    |
| <ٲ>   | U+06C1    | Heh Goal                            |               |     | ✓  |    | ✓  |    |    | ✓  |
| <ٲ>   | U+06C2    | Heh Goal with Hamza Above           |               |     | ✓  |    | ✓  |    |    | ✓  |
| <ٲ>   | U+06C3    | Teh Marbuta Goal                    |               |     |    |    | ✓  |    |    | ✓  |
| <ٲ>   | U+06C4    | Waw with Ring                       |               |     | ✓  |    |    |    |    |    |

Continued on next page

TABLE 22—Continued from previous page

| Char. | Codepoint | Character Name               | Language Tags |     |    |    |    |    |    |    |
|-------|-----------|------------------------------|---------------|-----|----|----|----|----|----|----|
|       |           |                              | az            | ckb | ks | ms | pa | sd | ug | ur |
| <ۆ>   | U+06C6    | Oe                           | ✓             | ✓   | ✓  |    |    |    | ✓  |    |
| <ۇ>   | U+06C7    | U                            | ✓             |     |    |    |    |    | ✓  |    |
| <ۈ>   | U+06C8    | Yu                           |               |     |    |    |    |    | ✓  |    |
| <ۋ>   | U+06CA    | Waw with Two Dots Above      |               | ✓   |    |    |    |    |    |    |
| <ۋ>   | U+06CB    | Ve                           |               |     |    |    |    |    | ✓  |    |
| <ۏ>   | U+06CC    | Farsi Yeh                    | ✓             | ✓   | ✓  |    | ✓  |    |    | ✓  |
| <ۏ>   | U+06CE    | Yeh with Small V             | ✓             | ✓   |    |    |    |    |    |    |
| <ۋ>   | U+06CF    | Waw with Dot Above           |               |     |    | ✓  |    |    |    |    |
| <ۏ>   | U+06D0    | E                            |               |     |    |    |    |    | ✓  |    |
| <ۏ>   | U+06D2    | Yeh Barree                   |               |     |    |    | ✓  |    |    | ✓  |
| <ۏ>   | U+06D3    | Yeh Barree with Hamza Above  |               |     |    |    | ✓  |    |    | ✓  |
| <ۏ>   | U+06D5    | Ae                           |               | ✓   |    |    |    |    | ✓  |    |
| <ۏ>   | U+06FD    | Sign Sindhi Ampersand        |               |     |    |    |    | ✓  |    |    |
| <ۏ>   | U+06FE    | Sign Sindhi Postposition Men |               |     |    |    |    | ✓  |    |    |
| <ۏ>   | U+0762    | Keheh with Dot Above         |               |     |    | ✓  |    |    |    |    |
| <ۏ>   | U+0762    | Keheh with Dot Above         |               |     |    | ✓  |    |    |    |    |
| <ۏ>   | U+0763    | Keheh with Three Dots Above  |               |     |    | ✓  |    |    |    |    |
| <ۏ>   | U+0768    | Noon with Small Tah Above    |               |     |    |    | ✓  |    |    |    |
| <ۏ>   | U+076C    | Reh with Hamza Above         |               |     | ✓  |    |    |    |    |    |
| <ۏ>   | U+08C7    | Lam with Small Tah Above     |               |     |    |    | ✓  |    |    |    |

## B. Language Model Experiments

### B.1. Character Language Models

Full character language model results are shown here for Kashmiri (Table 23), Kurdish Sorani (Table 24), Malay (Table 25), Punjabi Shahmukhi (Table 26), Sindhi (Table 27), South Azerbaijani (Table 28), Uyghur (Table 29), and Urdu (Table 30).

Assuming the significance level of  $\alpha = 0.05$ , the results for Kashmiri in Table 23, which is the smallest dataset, indicate that cross-entropy improvements for  $n$ -gram orders 4, 9, 10 are statistically significant. For the 8-grams, the MW and BW tests indicate borderline significance

TABLE 23. Significance tests for Kashmiri character  $n$ -gram language models

| $n$ | $\Delta_\mu$ |      | WS     |        |        |       | MW     |       | BM     |       |
|-----|--------------|------|--------|--------|--------|-------|--------|-------|--------|-------|
|     | $\Delta$     | %    | $L$    | $H$    | $t$    | $p$   | $t$    | $p$   | $t$    | $p$   |
| 3   | -0.003       | 0.08 | -0.015 | 0.009  | -0.458 | 0.647 | 5242.0 | 0.555 | -0.588 | 0.557 |
| 4   | -0.021       | 0.71 | -0.036 | -0.006 | -2.722 | 0.007 | 5946.0 | 0.021 | -2.356 | 0.019 |
| 5   | -0.006       | 0.21 | -0.024 | 0.013  | -0.607 | 0.544 | 5247.0 | 0.547 | -0.601 | 0.549 |
| 6   | -0.006       | 0.22 | -0.023 | 0.012  | -0.652 | 0.515 | 5327.0 | 0.425 | -0.793 | 0.429 |
| 7   | -0.005       | 0.2  | -0.024 | 0.013  | -0.578 | 0.564 | 5276.0 | 0.501 | -0.671 | 0.503 |
| 8   | -0.016       | 0.63 | -0.034 | 0.002  | -1.791 | 0.075 | 5802.0 | 0.05  | -1.977 | 0.05  |
| 9   | -0.022       | 0.85 | -0.04  | -0.004 | -2.468 | 0.014 | 6007.0 | 0.014 | -2.51  | 0.013 |
| 10  | -0.028       | 1.07 | -0.046 | -0.009 | -2.96  | 0.003 | 6065.0 | 0.009 | -2.676 | 0.008 |

TABLE 24. Significance tests for Kurdish (Sorani) character  $n$ -gram language models

| $n$ | $\Delta_\mu$ |      | WS     |        |         |     | MW     |     | BM       |     |
|-----|--------------|------|--------|--------|---------|-----|--------|-----|----------|-----|
|     | $\Delta$     | %    | $L$    | $H$    | $t$     | $p$ | $t$    | $p$ | $t$      | $p$ |
| 3   | -0.01        | 0.36 | -0.011 | -0.009 | -22.395 | 0.0 | 9916.0 | 0.0 | -115.778 | 0.0 |
| 4   | -0.003       | 0.16 | -0.004 | -0.002 | -6.687  | 0.0 | 7535.0 | 0.0 | -7.394   | 0.0 |
| 5   | -0.004       | 0.24 | -0.005 | -0.003 | -6.288  | 0.0 | 7369.0 | 0.0 | -6.724   | 0.0 |
| 6   | -0.003       | 0.2  | -0.004 | -0.002 | -4.633  | 0.0 | 6718.0 | 0.0 | -4.547   | 0.0 |
| 7   | -0.004       | 0.28 | -0.005 | -0.002 | -5.768  | 0.0 | 7270.0 | 0.0 | -6.294   | 0.0 |
| 8   | -0.006       | 0.44 | -0.007 | -0.004 | -8.596  | 0.0 | 8040.0 | 0.0 | -10.003  | 0.0 |
| 9   | -0.005       | 0.43 | -0.006 | -0.004 | -8.741  | 0.0 | 8006.0 | 0.0 | -9.734   | 0.0 |
| 10  | -0.005       | 0.39 | -0.006 | -0.003 | -7.075  | 0.0 | 7623.0 | 0.0 | -7.86    | 0.0 |

TABLE 25. Significance tests for Malay character  $n$ -gram language models

| $n$ | $\Delta_\mu$ |       | WS     |        |          |     | MW       |     | BM        |     |
|-----|--------------|-------|--------|--------|----------|-----|----------|-----|-----------|-----|
|     | $\Delta$     | %     | $L$    | $H$    | $t$      | $p$ | $t$      | $p$ | $t$       | $p$ |
| 3   | -0.065       | 1.818 | -0.066 | -0.064 | -194.951 | 0.0 | 10 000.0 | 0.0 | $-\infty$ | 0.0 |
| 4   | -0.062       | 2.036 | -0.063 | -0.061 | -143.242 | 0.0 | 10 000.0 | 0.0 | $-\infty$ | 0.0 |
| 5   | -0.06        | 2.232 | -0.061 | -0.059 | -135.144 | 0.0 | 10 000.0 | 0.0 | $-\infty$ | 0.0 |
| 6   | -0.064       | 2.526 | -0.065 | -0.063 | -132.604 | 0.0 | 10 000.0 | 0.0 | $-\infty$ | 0.0 |
| 7   | -0.067       | 2.736 | -0.068 | -0.066 | -135.678 | 0.0 | 10 000.0 | 0.0 | $-\infty$ | 0.0 |
| 8   | -0.07        | 2.885 | -0.071 | -0.069 | -126.796 | 0.0 | 10 000.0 | 0.0 | $-\infty$ | 0.0 |
| 9   | -0.07        | 2.931 | -0.071 | -0.069 | -116.01  | 0.0 | 10 000.0 | 0.0 | $-\infty$ | 0.0 |
| 10  | -0.07        | 2.922 | -0.071 | -0.068 | -103.192 | 0.0 | 10 000.0 | 0.0 | $-\infty$ | 0.0 |

TABLE 26. Significance tests for Punjabi (Shahmukhi) character  $n$ -gram language models

| $n$ | $\Delta_\mu$ |      | WS     |        |         |     | MW       |     | BM        |     |
|-----|--------------|------|--------|--------|---------|-----|----------|-----|-----------|-----|
|     | $\Delta$     | %    | $L$    | $H$    | $t$     | $p$ | $t$      | $p$ | $t$       | $p$ |
| 3   | -0.011       | 0.32 | -0.012 | -0.01  | -34.213 | 0.0 | 10 000.0 | 0.0 | $-\infty$ | 0.0 |
| 4   | -0.013       | 0.44 | -0.013 | -0.012 | -35.959 | 0.0 | 10 000.0 | 0.0 | $-\infty$ | 0.0 |
| 5   | -0.01        | 0.39 | -0.011 | -0.009 | -24.399 | 0.0 | 9941.0   | 0.0 | -167.411  | 0.0 |
| 6   | -0.008       | 0.34 | -0.009 | -0.007 | -16.158 | 0.0 | 9667.0   | 0.0 | -29.955   | 0.0 |
| 7   | -0.007       | 0.33 | -0.008 | -0.006 | -13.856 | 0.0 | 9429.0   | 0.0 | -21.268   | 0.0 |
| 8   | -0.006       | 0.3  | -0.008 | -0.004 | -6.491  | 0.0 | 8928.0   | 0.0 | -14.653   | 0.0 |
| 9   | -0.007       | 0.36 | -0.009 | -0.006 | -7.799  | 0.0 | 9061.0   | 0.0 | -16.395   | 0.0 |
| 10  | -0.007       | 0.34 | -0.009 | -0.005 | -6.11   | 0.0 | 8609.0   | 0.0 | -12.104   | 0.0 |

TABLE 27. Significance tests for Sindhi character  $n$ -gram language models

| $n$ | $\Delta_\mu$ |       | WS     |        |        |       | MW     |       | BM     |       |
|-----|--------------|-------|--------|--------|--------|-------|--------|-------|--------|-------|
|     | $\Delta$     | %     | $L$    | $H$    | $t$    | $p$   | $t$    | $p$   | $t$    | $p$   |
| 3   | 0.001        | -0.03 | -0.002 | 0.004  | 0.71   | 0.479 | 4649.0 | 0.392 | 0.853  | 0.395 |
| 4   | 0.011        | -0.28 | 0.005  | 0.016  | 3.628  | 0.0   | 3461.0 | 0.0   | 3.935  | 0.0   |
| 5   | -0.018       | 0.47  | -0.028 | -0.008 | -3.451 | 0.001 | 6718.0 | 0.0   | -4.51  | 0.0   |
| 6   | -0.028       | 0.79  | -0.039 | -0.018 | -5.183 | 0.0   | 7477.0 | 0.0   | -7.109 | 0.0   |
| 7   | -0.016       | 0.46  | -0.025 | -0.006 | -3.309 | 0.001 | 6714.0 | 0.0   | -4.448 | 0.0   |
| 8   | -0.024       | 0.71  | -0.033 | -0.014 | -4.775 | 0.0   | 6864.0 | 0.0   | -4.977 | 0.0   |
| 9   | -0.015       | 0.45  | -0.026 | -0.004 | -2.715 | 0.007 | 6119.0 | 0.006 | -2.814 | 0.005 |
| 10  | -0.014       | 0.42  | -0.026 | -0.003 | -2.443 | 0.015 | 5889.0 | 0.03  | -2.212 | 0.028 |

TABLE 28. Significance tests for South Azerbaijani character  $n$ -gram language models

| $n$ | $\Delta_\mu$ |       | WS     |        |         |       | MW     |       | BM       |       |
|-----|--------------|-------|--------|--------|---------|-------|--------|-------|----------|-------|
|     | $\Delta$     | %     | $L$    | $H$    | $t$     | $p$   | $t$    | $p$   | $t$      | $p$   |
| 3   | -0.012       | 0.418 | -0.013 | -0.011 | -24.371 | 0.0   | 9953.0 | 0.0   | -192.439 | 0.0   |
| 4   | -0.004       | 0.186 | -0.005 | -0.003 | -6.649  | 0.0   | 7482.0 | 0.0   | -7.222   | 0.0   |
| 5   | -0.002       | 0.131 | -0.003 | -0.001 | -3.133  | 0.002 | 6137.0 | 0.005 | -2.866   | 0.005 |
| 6   | -0.001       | 0.093 | -0.003 | 0.0    | -1.993  | 0.048 | 5837.0 | 0.041 | -2.072   | 0.04  |
| 7   | -0.002       | 0.14  | -0.003 | 0.0    | -2.526  | 0.012 | 5895.0 | 0.029 | -2.227   | 0.027 |
| 8   | -0.002       | 0.143 | -0.003 | 0.0    | -2.52   | 0.013 | 5997.0 | 0.015 | -2.489   | 0.014 |
| 9   | -0.002       | 0.13  | -0.003 | 0.0    | -2.348  | 0.02  | 5945.0 | 0.021 | -2.352   | 0.02  |
| 10  | -0.002       | 0.171 | -0.004 | -0.001 | -2.889  | 0.004 | 6238.0 | 0.002 | -3.107   | 0.002 |



TABLE 29. Significance tests for Uyghur character  $n$ -gram language models

| $n$ | $\Delta_\mu$ |       | WS     |        |        |       | MW     |      | BM     |       |
|-----|--------------|-------|--------|--------|--------|-------|--------|------|--------|-------|
|     | $\Delta$     | %     | $L$    | $H$    | $t$    | $p$   | $t$    | $p$  | $t$    | $p$   |
| 3   | -0.002       | 0.074 | -0.003 | -0.001 | -4.42  | 0.0   | 6636.0 | 0.0  | -4.273 | 0.0   |
| 4   | -0.001       | 0.051 | -0.002 | 0.0    | -2.257 | 0.025 | 5801.0 | 0.05 | -1.981 | 0.049 |
| 5   | -0.004       | 0.203 | -0.005 | -0.003 | -7.131 | 0.0   | 7690.0 | 0.0  | -7.964 | 0.0   |
| 6   | -0.004       | 0.219 | -0.005 | -0.003 | -6.315 | 0.0   | 7394.0 | 0.0  | -6.829 | 0.0   |
| 7   | -0.003       | 0.164 | -0.004 | -0.002 | -4.348 | 0.0   | 6661.0 | 0.0  | -4.353 | 0.0   |
| 8   | -0.004       | 0.225 | -0.005 | -0.002 | -5.126 | 0.0   | 6907.0 | 0.0  | -5.133 | 0.0   |
| 9   | -0.004       | 0.24  | -0.006 | -0.002 | -4.549 | 0.0   | 6699.0 | 0.0  | -4.476 | 0.0   |
| 10  | -0.005       | 0.295 | -0.006 | -0.003 | -6.446 | 0.0   | 7460.0 | 0.0  | -6.977 | 0.0   |

TABLE 30. Significance tests for Urdu character  $n$ -gram language models

| $n$ | $\Delta_\mu$ |      | WS     |        |         |       | MW     |     | BM      |     |
|-----|--------------|------|--------|--------|---------|-------|--------|-----|---------|-----|
|     | $\Delta$     | %    | $L$    | $H$    | $t$     | $p$   | $t$    | $p$ | $t$     | $p$ |
| 3   | -0.003       | 0.11 | -0.004 | -0.001 | -4.372  | 0.0   | 6646.0 | 0.0 | -4.306  | 0.0 |
| 4   | -0.005       | 0.22 | -0.005 | -0.004 | -17.256 | 0.0   | 9575.0 | 0.0 | -34.762 | 0.0 |
| 5   | -0.004       | 0.2  | -0.004 | -0.003 | -9.06   | 0.0   | 8813.0 | 0.0 | -15.737 | 0.0 |
| 6   | -0.004       | 0.26 | -0.006 | -0.003 | -5.352  | 0.0   | 8363.0 | 0.0 | -11.571 | 0.0 |
| 7   | -0.004       | 0.24 | -0.005 | -0.002 | -4.601  | 0.0   | 8246.0 | 0.0 | -10.74  | 0.0 |
| 8   | -0.002       | 0.14 | -0.004 | 0.0    | -2.35   | 0.02  | 8220.0 | 0.0 | -10.431 | 0.0 |
| 9   | -0.002       | 0.13 | -0.005 | 0.001  | -1.24   | 0.217 | 7896.0 | 0.0 | -8.207  | 0.0 |
| 10  | -0.005       | 0.37 | -0.008 | -0.002 | -3.422  | 0.001 | 8344.0 | 0.0 | -10.745 | 0.0 |

disagreeing with the WS test. The discrepancy observed between lower orders is probably due to overfitting as the dataset is tiny. Punjabi (Shahmukhi) and Kurdish (Sorani) are the second and third biggest datasets, respectively, and the results in Table 26 and Table 24 indicate statistically significant improvements across the board, with all the three tests agreeing with each other. Sindhi (Table 27) is similar in some respects to Kashmiri in that the low  $n$ -gram orders of 3 and 4 are not very reliable, while the results for the rest of the orders indicate significant improvements. While overfitting may play a certain role, upon informal inspection it appears that, similar to Kashmiri, the Sindhi dataset is quite noisy, even after filtering. Finally, the results for Urdu in Table 30 indicate that, similar to Kurdish (Sorani), Punjabi (Shahmukhi), South Azerbaijani and Uyghur the improvements are statistically significant across the board. The best results are obtained for Malay (Table 25) with up to 2.9% improvement in character entropy.

## B.2. Word Language Models

Full word language model results are shown here for Kashmiri (Table 31), Kurdish Sorani (Table 32), Malay (Table 33), Punjabi Shahmukhi (Table 34), Sindhi (Table 35), South Azerbaijani (Table 36), Uyghur (Table 37), and Urdu (Table 38).

As can be seen from Table 31, the hypothesis testing for Kashmiri shows that the null hypothesis is confirmed by all the three algorithms for all the orders  $n$  most of the time. This is evident from the tests'  $p$ -values (these exceed the significance level  $\alpha = 0.05$ ) as well as the  $t$ -test confidence intervals that contain the null hypothesis, which in our case corresponds to the zero difference in means. This is likely the artifact of the models overfitting a very small dataset, even for relatively less scarce bigrams. It is interesting to note that for the rest of the languages the alternative hypotheses for all the models is uniformly confirmed: the small decrease in cross-entropy expressed as bits per word observed for all the languages and configurations is statistically significant.

While the relative improvements for Kurdish Sorani (Table 32), Punjabi Shahmukhi (Table 34), South Azerbaijani (Table 36), Uyghur (Table 37) and Urdu (Table 38) are relatively tiny, possibly due to the relatively small number of modifications compared to the overall size of the datasets, the relative improvements to Sindhi models (Table 35) are over one percent for all the configurations. This may indeed be correlated with the highest number of per-word token modifications for Sindhi among all the languages, denoted  $R_w$  in Table 12. The best results are obtained for Malay (Table 33), with up to 3.5% improvement in word entropy.

## References

- Aazim, Muzaffar, Kamal Mansour, and Roozbeh Pournader (2009). *Proposal to add two Kashmiri characters and one annotation to the Arabic block*. Tech. rep. L2/09-176. Unicode Consortium.
- Abadi, Martín et al. (2016). "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." In: *CoRR* abs/1603.04467.
- Abdelali, Ahmed et al. (2014). "The AMARA Corpus: Building Parallel Language Resources for the Educational Domain." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 1856–1862.
- Abdullah, Farhanah et al. (2020). "Jawi Script and the Malay Society: Historical Background and Development." In: *International Journal of Management (IJM)* 11.7, 68–78.

TABLE 31. Significance tests for Kashmiri word  $n$ -gram language models

| $n$ | $\Delta_\mu$ |       | WS     |       |        |       | MW     |       | BM     |       |
|-----|--------------|-------|--------|-------|--------|-------|--------|-------|--------|-------|
|     | $\Delta$     | %     | $L$    | $H$   | $t$    | $p$   | $t$    | $p$   | $t$    | $p$   |
| 2   | 0.034        | -0.33 | -0.008 | 0.076 | 1.595  | 0.112 | 4301.0 | 0.088 | 1.72   | 0.087 |
| 3   | -0.039       | 0.39  | -0.088 | 0.011 | -1.531 | 0.127 | 5589.0 | 0.15  | -1.447 | 0.149 |
| 4   | -0.022       | 0.22  | -0.067 | 0.023 | -0.957 | 0.34  | 5308.0 | 0.452 | -0.751 | 0.454 |
| 5   | 0.017        | -0.17 | -0.029 | 0.063 | 0.727  | 0.468 | 4817.0 | 0.656 | 0.443  | 0.659 |

TABLE 32. Significance tests for Kurdish (Sorani) word  $n$ -gram language models

| $n$ | $\Delta_\mu$ |      | WS     |        |         |     | MW     |     | BM      |     |
|-----|--------------|------|--------|--------|---------|-----|--------|-----|---------|-----|
|     | $\Delta$     | %    | $L$    | $H$    | $t$     | $p$ | $t$    | $p$ | $t$     | $p$ |
| 2   | -0.031       | 0.41 | -0.038 | -0.024 | -8.678  | 0.0 | 8058.0 | 0.0 | -10.074 | 0.0 |
| 3   | -0.034       | 0.49 | -0.04  | -0.028 | -10.393 | 0.0 | 8492.0 | 0.0 | -13.09  | 0.0 |
| 4   | -0.034       | 0.49 | -0.041 | -0.027 | -9.297  | 0.0 | 8237.0 | 0.0 | -11.275 | 0.0 |
| 5   | -0.035       | 0.5  | -0.042 | -0.028 | -10.406 | 0.0 | 8464.0 | 0.0 | -13.209 | 0.0 |

TABLE 33. Significance tests for Malay word  $n$ -gram language models

| $n$ | $\Delta_\mu$ |       | WS     |        |          |     | MW       |     | BM        |     |
|-----|--------------|-------|--------|--------|----------|-----|----------|-----|-----------|-----|
|     | $\Delta$     | %     | $L$    | $H$    | $t$      | $p$ | $t$      | $p$ | $t$       | $p$ |
| 2   | -0.358       | 2.935 | -0.362 | -0.355 | -185.4   | 0.0 | 10 000.0 | 0.0 | $-\infty$ | 0.0 |
| 3   | -0.394       | 3.319 | -0.399 | -0.389 | -166.053 | 0.0 | 10 000.0 | 0.0 | $-\infty$ | 0.0 |
| 4   | -0.403       | 3.411 | -0.408 | -0.398 | -166.567 | 0.0 | 10 000.0 | 0.0 | $-\infty$ | 0.0 |
| 5   | -0.411       | 3.479 | -0.416 | -0.406 | -170.479 | 0.0 | 10 000.0 | 0.0 | $-\infty$ | 0.0 |

TABLE 34. Significance tests for Punjabi (Shahmukhi) word  $n$ -gram language models

| $n$ | $\Delta_\mu$ |      | WS     |        |         |     | MW     |     | BM      |     |
|-----|--------------|------|--------|--------|---------|-----|--------|-----|---------|-----|
|     | $\Delta$     | %    | $L$    | $H$    | $t$     | $p$ | $t$    | $p$ | $t$     | $p$ |
| 2   | -0.015       | 0.15 | -0.018 | -0.012 | -10.834 | 0.0 | 8633.0 | 0.0 | -14.295 | 0.0 |
| 3   | -0.017       | 0.19 | -0.021 | -0.014 | -9.338  | 0.0 | 8237.0 | 0.0 | -11.232 | 0.0 |
| 4   | -0.02        | 0.23 | -0.025 | -0.016 | -9.472  | 0.0 | 8318.0 | 0.0 | -11.75  | 0.0 |
| 5   | -0.02        | 0.23 | -0.024 | -0.016 | -8.916  | 0.0 | 8137.0 | 0.0 | -10.521 | 0.0 |

Afzal, Muhammad and Sarmad Hussain (2001). "Urdu computing standards: development of Urdu Zabta Takhti (UZT) 1.01." In: *Proceedings of IEEE International Multi Topic Conference (INMIC): Technology for the 21st Century*. IEEE. Lahore, Pakistan, 216–222.

TABLE 35. Significance tests for Sindhi word  $n$ -gram language models

| $n$ | $\Delta_\mu$ |      | WS     |        |         |     | MW     |     | BM       |     |
|-----|--------------|------|--------|--------|---------|-----|--------|-----|----------|-----|
|     | $\Delta$     | %    | $L$    | $H$    | $t$     | $p$ | $t$    | $p$ | $t$      | $p$ |
| 2   | -0.159       | 1.06 | -0.171 | -0.146 | -24.103 | 0.0 | 9948.0 | 0.0 | -180.99  | 0.0 |
| 3   | -0.169       | 1.21 | -0.186 | -0.152 | -19.861 | 0.0 | 9753.0 | 0.0 | -54.902  | 0.0 |
| 4   | -0.177       | 1.26 | -0.192 | -0.161 | -22.847 | 0.0 | 9889.0 | 0.0 | -88.059  | 0.0 |
| 5   | -0.184       | 1.32 | -0.2   | -0.168 | -22.57  | 0.0 | 9898.0 | 0.0 | -108.596 | 0.0 |

TABLE 36. Significance tests for South Azerbaijani word  $n$ -gram language models

| $n$ | $\Delta_\mu$ |       | WS     |        |        |     | MW     |     | BM      |     |
|-----|--------------|-------|--------|--------|--------|-----|--------|-----|---------|-----|
|     | $\Delta$     | %     | $L$    | $H$    | $t$    | $p$ | $t$    | $p$ | $t$     | $p$ |
| 2   | -0.031       | 0.374 | -0.038 | -0.024 | -9.016 | 0.0 | 8113.0 | 0.0 | -10.307 | 0.0 |
| 3   | -0.027       | 0.332 | -0.034 | -0.02  | -7.376 | 0.0 | 7655.0 | 0.0 | -7.913  | 0.0 |
| 4   | -0.031       | 0.397 | -0.038 | -0.025 | -9.711 | 0.0 | 8325.0 | 0.0 | -11.755 | 0.0 |
| 5   | -0.028       | 0.358 | -0.035 | -0.021 | -8.108 | 0.0 | 7797.0 | 0.0 | -8.828  | 0.0 |

TABLE 37. Significance tests for Uyghur word  $n$ -gram language models

| $n$ | $\Delta_\mu$ |       | WS     |        |        |     | MW     |     | BM     |     |
|-----|--------------|-------|--------|--------|--------|-----|--------|-----|--------|-----|
|     | $\Delta$     | %     | $L$    | $H$    | $t$    | $p$ | $t$    | $p$ | $t$    | $p$ |
| 2   | -0.024       | 0.2   | -0.033 | -0.015 | -5.14  | 0.0 | 6961.0 | 0.0 | -5.297 | 0.0 |
| 3   | -0.038       | 0.332 | -0.05  | -0.027 | -6.721 | 0.0 | 7537.0 | 0.0 | -7.397 | 0.0 |
| 4   | -0.034       | 0.294 | -0.045 | -0.022 | -5.689 | 0.0 | 7177.0 | 0.0 | -6.029 | 0.0 |
| 5   | -0.042       | 0.367 | -0.055 | -0.029 | -6.319 | 0.0 | 7381.0 | 0.0 | -6.774 | 0.0 |

TABLE 38. Significance tests for Urdu word  $n$ -gram language models

| $n$ | $\Delta_\mu$ |      | WS     |        |        |       | MW     |       | BM     |       |
|-----|--------------|------|--------|--------|--------|-------|--------|-------|--------|-------|
|     | $\Delta$     | %    | $L$    | $H$    | $t$    | $p$   | $t$    | $p$   | $t$    | $p$   |
| 2   | -0.006       | 0.08 | -0.008 | -0.003 | -4.923 | 0.0   | 6796.0 | 0.0   | -4.804 | 0.0   |
| 3   | -0.004       | 0.06 | -0.007 | -0.002 | -3.507 | 0.001 | 6394.0 | 0.001 | -3.531 | 0.001 |
| 4   | -0.006       | 0.1  | -0.009 | -0.003 | -4.2   | 0.0   | 6637.0 | 0.0   | -4.285 | 0.0   |
| 5   | -0.007       | 0.1  | -0.009 | -0.004 | -4.927 | 0.0   | 6842.0 | 0.0   | -4.914 | 0.0   |

Ahmad, Humza and Laszlo Erdodi (2021). “Overview of phishing landscape and homographs in Arabic domain names.” In: *Security and Privacy* 4.4, 1–14.

Ahmadi, Sina (2019). “A rule-based Kurdish text transliteration system.” In: *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 18.2, 1–8.

- Anastasopoulos, Antonios et al. (2020). "TICO-19: the Translation Initiative for COVID-19." In: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Online: Association for Computational Linguistics.
- Anuvaad, Team (2022). *Project Anuvaad: Parallel Corpus*. Accessed: June, 2022; <https://anuvaaad.org/>.
- Aqtay, Gulayhan (2020). "The New Kazakh Alphabet Based on Latin Script." In: *Language and Society in Kazakhstan: The Kazakh Context*. Ed. by Gulayhan Aqtay and Cem Erdem. Vol. 6. Turkic Studies. Poznań, Poland: Wydawnictwo Naukowe, Adam Mickiewicz University, 23–36.
- Aronson, Howard (1996). "Yiddish." In: *The World's Writing Systems*. Ed. by Peter Daniels and William Bright. Oxford: Oxford University Press. Chap. 61, 735–742.
- Avanzini, Alessandra (2009). "Origin and classification of the Ancient South Arabian languages." In: *Journal of Semitic Studies* 54.1, 205–220.
- Bahdanau, Dmitry, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural machine translation by jointly learning to align and translate." In: *Proceedings of 3rd International Conference on Learning Representations (ICLR)*. San Diego, CA.
- Bashir, Elena (2015). "The Brahui Language: Recovering the Past, Documenting the Present, and Pondering the Future." In: *Proceedings of International Conference on Brahui Language and Culture*. Brahui Academy, Baluchistan, Pakistan. Islamabad, Pakistan, 1–27.
- Bashir, Elena and Thomas J. Connors (2019). *A Descriptive Grammar of Hindko, Panjabi, and Saraiki*. Vol. 4. Mouton-CASL Grammar Series (MCASL). De Gruyter Mouton.
- Bashir, Elena, Sarmad Hussain, and Deborah Anderson (2006). *Proposal for characters for Khowar, Torwali, and Burushaski*. Tech. rep. L2-06/149. Unicode Consortium.
- Bauer, Thomas (1996). "Arabic Writing." In: *The World's Writing Systems*. Ed. by Peter Daniels and William Bright. Oxford: Oxford University Press. Chap. 50, 559–563.
- Bhardwaj, Mangat Rai (2016). *Panjabi: A Comprehensive Grammar*. Routledge Comprehensive Grammars. Routledge.
- Bhatt, Rakesh M. (2015). "Script Choice, Language Loss and the Politics of Anamnesis: Kashmiri in Diaspora." In: *Language, Literacy and Diversity*. Ed. by Christopher Stroud and Mastin Prinsloo. Routledge Critical Studies in Multilingualism. Routledge, 130–147.
- Bosworth, Clifford Edmund (2011). "Arrān." In: *Encyclopædia Iranica*. Ed. by Ehsan Yarshater. Vol. II/5. Online. Leiden, The Netherlands: Brill, 520–522.
- Boudelaa, Sami, Manuel Perea, and Manuel Carreiras (2020). "Matrices of the frequency and similarity of Arabic letters and allographs." In: *Behavior Research Methods* 52.5, 1893–1905.

- Bozarslan, Hamit, Cengiz Gunes, and Veli Yadirgi (2021). "Kurdish Language." In: *The Cambridge History of the Kurds*. Part V. Cambridge, UK: Cambridge University Press, 601–684.
- Braginsky, Vladimir I. (1975). "Some remarks on the structure of the "Sya'ir Perahu" by Hamzah Fansuri." In: *Bijdragen tot de Taal-, Land- en Volkenkunde / Journal of the Humanities and Social Sciences of Southeast Asia and Oceania* 131.4, 407–426.
- Bright, William (1999). "A matter of typology: Alphasyllabaries and abugidas." In: *Written Language & Literacy* 2.1, 45–55.
- Brose, Michael C. (2017). "The Medieval Uyghurs of the 8th through 14th Centuries." In: *Oxford Research Encyclopedia of Asian History*. Online. Oxford University Press, 1–20.
- Brown, Peter F. et al. (1992). "An estimate of an upper bound for the entropy of English." In: *Computational Linguistics* 18.1, 31–40.
- Brunner, Edgar and Ullrich Munzel (2000). "The nonparametric Behrens-Fisher problem: Asymptotic theory and small-sample approximation." In: *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 42.1, 17–25.
- Buljina, Harun (2019). "Empire, Nation, and the Islamic World: Bosnian Muslim Reformists between the Habsburg and Ottoman Empires, 1901–1914." PhD thesis. New York: Columbia University.
- Campbell, George L. (1994). "Kurdish." In: *Concise Compendium of the World's Languages*. Ed. by George L. Campbell. London: Routledge. Chap. 5, 288–292.
- Castilla, Nuria de (2019). "Uses and Written Practices in Aljamiado Manuscripts." In: *Creating Standards: Interactions with Arabic script in 12 manuscript cultures*. Ed. by Dmitry Bondarev, Alessandro Gori, and Lameen Souag. Vol. 16. Studies in Manuscript Cultures. De Gruyter.
- Caswell, Isaac et al. (2020). "Language ID in the Wild: Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 6588–6608.
- Chitralli, Rehmat Aziz Khan (2020a). *Proposal to include Indus Kobistani Language alphabets*. Tech. rep. L2/20-157. Unicode Consortium.
- (2020b). *Proposal to include Kalasba Language alphabets*. Tech. rep. L2/20-091. Unicode Consortium.
- Coluzzi, Paolo (2020). "Jawi, an endangered orthography in the Malaysian linguistic landscape." In: *International Journal of Multilingualism*, 1–17.
- Conneau, Alexis et al. (2020). "Unsupervised Cross-lingual Representation Learning at Scale." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 8440–8451.

- Cotterell, Ryan et al. (2018). "Are All Languages Equally Hard to Language-Model?" In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 536–541.
- Coulmas, Florian (1999). *The Blackwell Encyclopedia of Writing Systems*. Oxford, UK: John Wiley & Sons.
- Daniels, Peter T. (2013). "The Arabic writing system." In: *The Oxford Handbook of Arabic Linguistics*. Ed. by Jonathan Owens. Oxford, UK: Oxford University Press. Chap. 18, 422–431.
- (2014). "The type and spread of Arabic script." In: *The Arabic Script in Africa*. Ed. by Meikal Mumin and Kees Versteegh. Vol. 71. Studies in Semitic Languages and Linguistics. Brill, 25–39.
- Datta, A. K. (1984). "A Generalized Formal Approach for Description and Analysis of Major Indian Scripts." In: *IETE Journal of Research* 30.6, 155–161.
- Davis, Derek (2015). "Premchand plays chess." In: *Journal of the Royal Asiatic Society* 25.2, 269–300.
- DBP (2006). *Daftar kata bahasa Melayu: Rumi-Sebutan-Jawi [Malay Pronunciation Dictionary (Rumi-Jawi)]*. 5th. Kuala Lumpur, Malaysia: Dewan Bahasa Pustaka (DBP) [Institute of Language and Literature].
- Dillon, Michael (2009). *Xinjiang — China's Muslim Far Northwest*. Durham East Asia Series. London and New York: Routledge.
- Dow, Hugh (1976). "A note on the Sindhi alphabet." In: *Asian Affairs* 7.1, 54–56.
- Dwyer, Arienne M. (2005). *The Xinjiang conflict: Uyghur identity, language policy, and political discourse*. Vol. 15. Policy Studies. Washington, D.C.: East-West Center Washington.
- Elsayed, Yahia and Ahmed Shosha (2018). "Large scale detection of IDN domain name masquerading." In: *Proceedings of 2018 APWG Symposium on Electronic Crime Research (eCrime)*. IEEE. San Diego, CA, 1–11.
- Esmaili, Kyumars Sheykh et al. (2013). "Building a test collection for Sorani Kurdish." In: *Proceedings of ACS International Conference on Computer Systems and Applications (AICCSA)*. IEEE. Ifrane, Morocco, 1–7.
- Evans, Lorna Priest and Andy Warren-Rothlin (2018). *Proposal to encode additional Arabic script characters for Hausa to the UCS*. Tech. rep. L2/18-094. Unicode Consortium.
- Eviatar, Zohar and Raphiq Ibrahim (2014). "Why is it Hard to Read Arabic?" In: *Handbook of Arabic Literacy: Insights and Perspectives*. Ed. by Elinor Saiegh-Haddad and R. Malatesha Joshi. Vol. 9. Literacy Studies. Springer, 77–96.
- Fan, Angela et al. (2021). "Beyond English-centric multilingual machine translation." In: *Journal of Machine Learning Research* 22.107, 1–48.

- Fedorova, Liudmila L (2012). "The development of structural characteristics of Brahmi script in derivative writing systems." In: *Written Language & Literacy* 15.1, 1–25.
- Friedmann, Naama and Manar Haddad-Hanna (2014). "Types of developmental dyslexia in Arabic." In: *Handbook of Arabic Literacy: Insights and Perspectives*. Ed. by Elinor Saiegh-Haddad and R. Malatesha Joshi. Vol. 9. Literacy Studies. Springer, 119–151.
- Gorman, Kyle (2016). "Pynini: A Python library for weighted finite-state grammar compilation." In: *Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata*. Berlin, Germany: Association for Computational Linguistics, 75–80.
- Gorman, Kyle and Richard Sproat (2021). *Finite-State Text Processing*. Synthesis Lectures on Human Language Technologies 50. Morgan & Claypool Publishers.
- Gowda, Thamme et al. (2021). "Many-to-English Machine Translation Tools, Data, and Pretrained Models." In: *arXiv preprint arXiv:2104.00290*.
- Gowda, Thamme et al. (2021). "Many-to-English Machine Translation Tools, Data, and Pretrained Models." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, 306–316.
- Grewal, Jagtar Singh (2004). "Historical Geography of the Punjab." In: *Journal of Punjab Studies* 11.1. Center of Sikh and Punjab Studies, UC Santa Barbara, 1–18.
- Gruendler, Beatrice (1993). *The development of the Arabic scripts: From the Nabatean era to the first Islamic century according to dated texts*. Vol. 43. Harvard Semitic Studies. Atlanta, Georgia: Scholars Press.
- Gutkin, Alexander, Cibu Johny, Raïmond Doctor, Brian Roark, et al. (2022). "Beyond Arabic: Software for Perso-Arabic Script Manipulation." In: *Proceedings of the 7th Arabic Natural Language Processing Workshop (WANLP2022)*. To appear. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.
- Gutkin, Alexander, Cibu Johny, Raïmond Doctor, Lawrence Wolf-Sonkin, et al. (2022). "Extensions to Brahmic script processing within the Nisaba library: new scripts, languages and utilities." In: *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 6450–6460.
- Haig, Geoffrey (2018). "The Iranian languages of northern Iraq." In: *The Languages and Linguistics of Western Asia: An Areal Perspective*. Ed. by Geoffrey Haig and Geoffrey Khan. Vol. 6. The World of Linguistics (WOL). Berlin, Munich & Boston: Walter de Gruyter. Chap. 3.3, 267–304.



- Hamut, Bahargül and Agnieszka Joniak-Lüthi (2015). "The language choices and script debates among the Uyghur in Xinjiang Uyghur Autonomous Region, China." In: *Linguistik Online* 70.1, 111–124.
- Haralambous, Yannis (2021). "Breaking Arabic: the creative inventiveness of Uyghur script reforms." In: *Design Regression*.
- Hatcher, Lynley (2008). "Script change in Azerbaijan: Acts of identity." In: *International Journal of the Sociology of Language* 192, 105–116.
- Heafield, Kenneth (2011). "KenLM: Faster and Smaller Language Model Queries." In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, 187–197.
- Heafield, Kenneth et al. (2013a). "Scalable Modified Kneser-Ney Language Model Estimation." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, 690–696.
- (2013b). "Scalable Modified Kneser-Ney Language Model Estimation." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Sofia, Bulgaria: Association for Computational Linguistics, 690–696.
- Hermena, Ehab W. and Erik D. Reichle (2020). "Insights from the study of Arabic reading." In: *Language and Linguistics Compass* 14.10, 1–26.
- Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory." In: *Neural Computation* 9.8, 1735–1780.
- Hussain, Sarmad et al. (2016). "Enabling multilingual domain names: addressing the challenges of the Arabic script top-level domains." In: *Journal of Cyber Policy* 1.1, 107–129.
- Hutchins, William John (1986). *Machine Translation: Past, Present, Future*. Ellis Horwood Series in Engineering Science. Chichester (West Sussex), UK: Ellis Horwood Chichester.
- ICANN (2011). *Arabic Case Study Team: Arabic Case Study Team Issues Report*. Internationalized Domain Names (IDN) Variant Issues Project. Internet Corporation for Assigned Names and Numbers (ICANN).
- (2015). *Task Force on Arabic Script IDN (TF-AIDN): Proposal for Arabic Script Root Zone LGR*. ICANN Internationalized Domain Names (IDN) Program: Proposal Documentation. Version 2.7. Internet Corporation for Assigned Names and Numbers (ICANN).
- ISO (1984). *ISO 233:1984: Transliteration of Arabic characters into Latin characters*. <https://www.iso.org/standard/4117.html>. International Organization for Standardization.
- (1993). *ISO 233-2:1993: Transliteration of Arabic characters into Latin characters — Part 2: Arabic language — Simplified transliteration*. <https://www.iso.org/standard/4118.html>. International Organization for Standardization.
- (1999). *ISO 233-3:1999: Transliteration of Arabic characters into Latin characters — Part 3: Persian language — Simplified transliteration*. <https://www.iso.org/standard/4119.html>. International Organization for Standardization.

- ://www.iso.org/standard/4118.html. International Organization for Standardization.
- Iyengar, Arvind (2018). "Variation in Perso-Arabic and Devanāgarī Sindhī orthographies: An overview." In: *Written Language & Literacy* 21.2, 169–197.
- Jahani, Carina and Agnes Korn (2013). "Balochi." In: *The Iranian Languages*. Ed. by Gernot Windfuhr. Routledge Language Family Series. Routledge, 710–768.
- Johanson, Lars and Éva Ágnes Csató (2021). *The Turkic Languages*. 2nd. Routledge Language Family Series. Routledge.
- Johny, Cibu et al. (2021). "Finite-state script normalization and processing utilities: The Nisaba Brahmic library." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, 14–23.
- Kachru, Braj B. (2016). "Kashmiri and other Dardic languages." In: *Linguistics in South Asia*. Ed. by Murray B. Emeneau and Charles A. Ferguson. Vol. 5. Current Trends in Linguistics. De Gruyter Mouton, 284–306.
- Kachru, Raj (2021). *Kashmiri Proverbs*. Chennai, Tamil Nadu, India: Notion Press Media.
- Kaplony, Andreas (2008). "What are those few dots for? Thoughts on the orthography of the Qurra Papyri (709–710), the Khurasan Parchments (755–777) and the inscription of the Jerusalem Dome of the Rock (692)." In: *Arabica* 55.Fasc. 1, 91–112.
- Kaye, Alan S. (1996). "Adaptations of Arabic Script." In: *The World's Writing Systems*. Ed. by Peter Daniels and William Bright. Oxford: Oxford University Press. Chap. 62, 743–762.
- Khalid, Hewa Salam (2015). "Kurdish dialect continuum, as a standardization solution." In: *International Journal of Kurdish Studies* 1.1, 27–39.
- Khaw, Nasha Bin Rodziadi (2015). "Study and Analysis of the Proto-Shāradā and Shāradā inscriptions in the Lahore Museum (Pakistan)." In: *Gandbaran Studies* 9, 87–114.
- King, Robert D. (2001). "The poisonous potency of script: Hindi and Urdu." In: *International Journal of the Sociology of Language* 150, 43–59.
- El-Kishky, Ahmed et al. (2021). "XLent: Mining a Large Cross-lingual Entity Dataset with Lexical-Semantic-Phonetic Word Alignment." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 10424–10430.
- Klein, Guillaume et al. (2017). "OpenNMT: Open-Source Toolkit for Neural Machine Translation." In: *Proceedings of ACL 2017, System Demonstrations*. Vancouver, Canada: Association for Computational Linguistics, 67–72.

- Koehn, Philipp (2004). "Statistical Significance Tests for Machine Translation Evaluation." In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain: Association for Computational Linguistics, 388–395.
- Kotzé, Ernst (2012). "Arabic Afrikaans – early standardisation of Afrikaans orthography: A discussion of The Afrikaans of the Cape Muslims by Achmat Davids." In: *Southern African Linguistics and Applied Language Studies* 30.3, 413–427.
- Koul, Omkar Nath (2006). *A Dictionary of Kashmiri Proverbs*. Educa Books.
- Koul, Omkar Nath and Kashi Wali (2004). *Modern Kashmiri Grammar*. Springfield, VA, USA: Dunwoody Press.
- (2015). *Kashmiri*. Vol. 03. LINCOM Grammar Handbooks (LGH). Munich, Germany: LINCOM Academic Publishers.
- Kratz, E. Ulrich (2002). "Jawi spelling and orthography: a brief review." In: *Indonesia and the Malay World* 30.86, 21–26.
- Kurzon, Dennis (2013). "Diacritics and the Perso-Arabic script." In: *Writing Systems Research* 5.2, 234–243.
- Lajwani, Ali Murad and Abdul Jaleel Mirjat (2021). "The Mystical Philosophy of Shah Abdul Latif Bhittai: A Study of Shah-Jo-Risalo." In: *Al-Hikmat: A Journal of Philosophy* 41, 61–71.
- Lekhwani, K. and B. Lekhwani (2014). *Sindhi Word Forms*. Raipur, India: Chattisgarh Sindhi Sahitya Sansthan (Akademi) [Chattisgarh Sindhi Academy of Letters].
- Lelyveld, David (1994). "Zuban-e Urdu-e Mu'alla and the Idol of Linguistic Origins." In: *Annual of Urdu Studies* 9.
- Liljegren, Henrik (2016). *A grammar of Palula*. Vol. 8. Studies in Diversity Linguistics. Berlin, Germany: Language Science Press.
- (2018). "Supporting and sustaining language vitality in Northern Pakistan." In: *The Routledge Handbook of Language Revitalization*. Ed. by Leanne Hinton, Leena Huss, and Gerald Roche. Routledge, 427–437.
- Luong, Thang, Hieu Pham, and Christopher D. Manning (2015). "Effective Approaches to Attention-based Neural Machine Translation." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 1412–1421.
- Mann, Henry B. and Donald R. Whitney (1947). "On a test of whether one of two random variables is stochastically larger than the other." In: *The Annals of Mathematical Statistics* 18.1, 50–60.
- Milo, Thomas (2002). "Authentic Arabic: A case study. right-to-left font structure, font design, and typography." In: *Manuscripta Orientalia* 8.1, 49–61.
- Mnih, Volodymyr et al. (2014). "Recurrent models of visual attention." In: *Proceedings of Neural Information Processing Systems (NIPS)*. Montreal, Canada, 2204–2212.

- Mohamed, Sayed (1968). *The Value of Dakbni Language and Literature (Special Lectures)*. Mysore, India: University of Mysore.
- Mohri, Mehryar (1996). "On some applications of finite-state automata theory to natural language processing." In: *Natural Language Engineering* 2.1, 61–80.
- (2009). "Weighted Automata Algorithms." In: *Handbook of Weighted Automata*. Ed. by Manfred Droste, Werner Kuich, and Heiko Vogler. Monographs in Theoretical Computer Science. Springer, 213–254.
- Mokari, Payam Ghaffarvand and Stefan Werner (2017). "Azerbaijani." In: *Journal of the International Phonetic Association* 47.2, 207–212.
- MS (2012). *Information Technology — Jawi Coded Character Set For Information Exchange*. Malaysian Standard MS 2443:2012. Department of Standards Malaysia, Ministry of International Trade and Industry (MITI).
- Muller, Benjamin et al. (2021). "When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 448–462.
- Mumin, Meikal (2014). "The Arabic Script in Africa: Understudied Literacy." In: *The Arabic Script in Africa*. Ed. by Meikal Mumin and Kees Versteegh. Vol. 71. Studies in Semitic Languages and Linguistics. Leiden, The Netherlands: Brill, 41–76.
- Naim, C. Mohammed (1971). "Arabic orthography and some non-Semitic languages." In: *Islam and its Cultural Divergence: Studies in Honor of Gustave E. von Grunebaum*. Ed. by Girdhari L. Tikku. Urbana, IL: University of Illinois Press, 113–144.
- Nemeth, Titus (2017). *Arabic Type-Making in the Machine Age: The Influence of Technology on the Form of Arabic Type, 1908–1993*. Vol. 14. Islamic Manuscripts and Books. Leiden, The Netherlands: Brill.
- Ngom, Fallou (2010). "Ajami scripts in the Senegalese speech community." In: *Journal of Arabic and Islamic Studies* 10, 1–23.
- Ngom, Fallou and Mustapha H. Kurfi (2017). "Ajamization of Islam in Africa." In: *Islamic Africa* 8.1-2, 1–12.
- Papineni, Kishore et al. (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation." In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 311–318.
- Parhami, Behrooz (2020). "Computers and Challenges of Writing in Persian: Explorations at the Intersection of Culture and Technology." In: *Visible Language* 54.1-2, 187–223.
- Patel, Neil, Charles Riley, and Jesus MacLean (2019). *Proposal to add Arabic letter JEEM WITH THREE DOTS ABOVE and JEEM WITH THREE DOTS BELOW*. Tech. rep. L2/19-118. Unicode Consortium.

- Paullada, Amandalynne et al. (2021). "Data and its (dis)contents: A survey of dataset development and use in machine learning research." In: *Patterns* 2.11, p. 100336.
- Ponti, Edoardo Maria et al. (2019). "Towards Zero-shot Language Modeling." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2900–2910.
- Popović, Maja (2016). "chrF deconstructed: beta parameters and n-gram weights." In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, 499–504.
- Post, Matt (2018). "A Call for Clarity in Reporting BLEU Scores." In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, 186–191.
- Post, Matt, Chris Callison-Burch, and Miles Osborne (2012). "Constructing Parallel Corpora for Six Indian Languages via Crowdsourcing." In: *Proceedings of the Seventh Workshop on Statistical Machine Translation*. Montréal, Canada: Association for Computational Linguistics, 401–409.
- Pournader, Roozbeh (2010). *Of hamza and other harakat*. Tech. rep. L2/10-455. Unicode Consortium.
- Prentice, D. J. (1990). "Malay (Indonesian and Malaysian)." In: *The Major Languages of East and South-East Asia*. Ed. by Bernard Comrie. London, UK: Routledge. Chap. 10, 185–207.
- Qutbuddin, Tahera (2007). "Arabic in India: A survey and classification of its uses, compared with Persian." In: *Journal of the American Oriental Society* 127.3, 315–338.
- Reimers, Nils and Iryna Gurevych (2020). "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 4512–4525.
- Ricci, Ronit (2011). *Islam translated: Literature, conversion, and the Arabic cosmopolis of South and Southeast Asia*. South Asia Across the Disciplines. Chicago, IL, USA: University of Chicago Press.
- Riezler, Stefan and John T. Maxwell (2005). "On Some Pitfalls in Automatic Evaluation and Significance Testing for MT." In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, 57–64.
- Rosenfeld, Ronald (2000). "Two decades of statistical language modeling: where do we go from here?" In: *Proceedings of the IEEE* 88.8, 1270–1278.

- Sachau, Edward C. (1910). *Alberuni's India: An Account of the Religion, Philosophy, Literature, Geography, Chronology, Astronomy, Customs, Laws and Astrology of India about A.D. 1030*. London: Kegan Paul, Trench, Trübner & Co.
- Satterthwaite, Franklin E. (1946). "An approximate distribution of estimates of variance components." In: *Biometrics Bulletin* 2.6, 110–114.
- Schmidt, Ruth Laila (2007). "Urdu." In: *The Indo-Aryan Languages*. Ed. by George Cardona and Dhanesh Jain. Routledge Language Family Series. Routledge, 315–385.
- Schuster, Mike and Kuldip K. Paliwal (1997). "Bidirectional Recurrent Neural Networks." In: *IEEE Transactions on Signal Processing* 45.11, 2673–2681.
- Schwenk, Holger et al. (2021). "CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 6490–6500.
- Sh., Rikza F. (2022). *Proposal to encode three Quranic Arabic characters*. Tech. rep. L2-22/153. Unicode Consortium.
- Shannon, Claude E. (1951). "Prediction and Entropy of Printed English." In: *The Bell System Technical Journal* 30.1, 50–64.
- Singh, Surinder and Ishwar Dayal Gaur (2009). *Sufism in Punjab: Mystics, Literature, and Shrines*. Delhi, India: Aakar Books.
- Snover, Matthew et al. (2006). "A Study of Translation Edit Rate with Targeted Human Annotation." In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, 223–231.
- Sproat, Richard (2003). "A formal computational analysis of Indic scripts." In: *In International Symposium on Indic Scripts: Past and Future*. Tokyo, Japan.
- Stahlberg, Felix (2020). "Neural machine translation: A review." In: *Journal of Artificial Intelligence Research* 69, 343–418.
- Suutarinen, Mikko (2013). "Arabic Script among China's Muslims: A Dongxiang folk story." In: *Studia Orientalia Electronica*. Ed. by Tiina Hyytiäinen et al. Vol. 113. Helsinki, Finland: Finnish Oriental Society, WS Bookwell Oy, 197–208.
- Thackston, Wheeler McIntosh (2006). *Sorani Kurdish: A Reference Grammar with Selected Readings*. The original link has disappeared.
- Tiedemann, Jörg (2012). "Parallel Data, Tools and Interfaces in OPUS." In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), 2214–2218.
- Torwali, Zubair (2019). "Early Writing in Torwali in Pakistan." In: *Teaching Writing to Children in Indigenous Languages: Instructional Practices from*

- Global Contexts*. Ed. by Ari Sherris and Joy Kreeft Peyton. New York: Routledge, 44–70.
- Uddin, Naeem and Jalal Uddin (2019). “A step towards Torwali machine translation: an analysis of morphosyntactic challenges in a low-resource language.” In: *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*. Dublin, Ireland: European Association for Machine Translation, 6–10.
- Unicode Consortium (2021). “Arabic.” In: *The Unicode Standard (Version 14.0.0)*. Mountain View, CA: Unicode Consortium. Chap. 9.2, 373–398.
- Versteegh, Kees (2001). “Arabic in Madagascar.” In: *Bulletin of the School of Oriental and African Studies* 64.2, 177–187.
- Welch, Bernard L. (1947). “The generalization of “Student’s” problem when several different population variances are involved.” In: *Biometrika* 34.1/2, 28–35.
- Wenzek, Guillaume et al. (2021). “Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation.” In: *Proceedings of the Sixth Conference on Machine Translation*. Online: Association for Computational Linguistics, 89–99.
- Whistler, Ken (2021). *Unicode Normalization Forms*. Tech. rep. TR15-51. Version 14.0.0. Unicode Consortium.
- Wilkens, Jens (2016). “Buddhism in the West Uyghur Kingdom and Beyond.” In: *Transfer of Buddhism across Central Asian networks (7th to 13th centuries)*. Ed. by Carmen Meinert. Vol. 8. Dynamics in the History of Religions. Brill. Chap. 6, 189–249.
- Wink, André (1991). *Al-Hind: The Making of the Indo-Islamic World: Early Medieval India and the Expansion of Islam 7th–11th Centuries*. Vol. 1. Leiden & New York: Brill.
- Winstedt, Richard Olaf (1961). “Malay Chronicles from Sumatra and Malaya.” In: *Historians of South-East Asia*. Ed. by D. G. E. Hall. Vol. II. Historical Writing on the Peoples of Asia. Oxford, UK: Oxford University Press, 24–28.
- Xue, Linting, Aditya Barua, et al. (2022). “ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models.” In: *Transactions of the Association for Computational Linguistics* 10, 291–306.
- Xue, Linting, Noah Constant, et al. (2021). “mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer.” In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 483–498.
- Yakup, Mahire et al. (2015). “Extending models of visual-word recognition to semicursive scripts: Evidence from masked priming in Uyghur.” In: *Journal of Experimental Psychology: Human Perception and Performance* 41.6, 1553–1562.

- Yarshater, Ehsan (2011). "The Iranian Language of Azerbaijan." In: *Encyclopædia Iranica*. Ed. by Ehsan Yarshater. Vol. III/3. Online. Leiden, The Netherlands: Brill, 238–245.
- Yassin, Rana, David L. Share, and Yasmin Shalhoub-Awwad (2020). "Learning to spell in Arabic: The impact of script-specific visual-orthographic features." In: *Frontiers in Psychology* 11, p. 2059.
- Yattoo, Altaf Hussain (2012). *The Islamization of Kashmir (A Study of Muslim Missionaries)*. Srinagar, Jammu and Kashmir, India: Gulshan Books.
- Zabell, Sandy L. (2008). "On Student's 1908 Article "The Probable Error of a Mean"." In: *Journal of the American Statistical Association* 103.481, 1–7.
- Zhang, Biao et al. (2020). "Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 1628–1639.



# Semitic Writings and Short Vowels

## Alternative Hypotheses in a Renewed View of The *Analytics of writing*

Joseph Dichy

*Abstract.* The fact that Semitic writings do not note short vowels in the body of words is an ancient question, which has been repeatedly asked in works on the history of writing, from Marcel Cohen's, James Février's, I.J. Gelb's... onwards, as well as in the field of Semitic studies (G.R. Driver, D. Diringier...).

A first answer—still repeatedly reproduced—mentions the weight of tradition, and assumes that vowels were just not conceptualised, which resulted in what was described, in a factually improper wording, as “consonantal writings”. The latter have been considered as a step in the “history of writing”, i.e., in the “development of the invention of writing” (M. Cohen), the ultimate result of which would have been the ancient Greek alphabet. Such a view can no longer be held, because:

(1) Semitic writings had noted long vowels at the end, then in the body of words from the 13th century B.C. onwards, and

(2) the now prevalent idea is that every writing system is related to the language and culture in which it emerges and develops.


Another ancient hypothesis (Février; M. Cohen), widely taken up by Arab linguists, suggests that Semitic morphological patterns make up for the lack of vowels. We show here that this hypothesis, which only covers a percentage of words, cannot be retained either.

Another hypothesis (Dichy 2017) is discussed: short vowels being subject to dialectal variation, their omission in standard script may result, in a form of partially ‘robust’ writing, featuring a level of abstraction (Vendryes, 1923) that allows it to be shared by a variety of dialects or a-kin languages. This hypothesis should not be considered teleologically: it is a consequence of writing structures, and not a feature Semitic writing systems could have been “devised for”. In assumes in addition that the writing system is a writing-to-sounds process, which is a mistaken view of reading.

General alternative hypotheses are summed up in a renewed conceptual frame. They have been developed in Dichy (1990, 2017, 2019). The general concept is that of the *Analytics of writing*. According to it, the emergence of a writing system stems from the way in which a given culture analyses the structures of its own language in a way that produces:

- finite inventories of phono-graphic units, i.e., of grapheme-segments (or letters), defined through intuitive phonological processes, in relation to the fact

---

Joseph Dichy  0000-0002-9123-7358  
Professor at the Canadian University Dubai  
E-mail: joseph.dichy@yahoo.fr

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 377–390. <https://doi.org/10.36824/2022-graf-dich>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

- that, in Semitic languages (as opposed to Indo-European ones) all syllables start with a consonant, and roots are exclusively consonantal;
- a projection of these letters on a graphic system, which becomes – once instituted—a conventional way of writing;
  - word-form structures, which play a central part in the reading-for-meaning process.

## 1. Introduction

Another way of putting the question included in the title would be:

- “How could a renewed theoretical approach of writing offer an adequate description of Semitic graphic systems?” and conversely:
- “To what extent does the analysis of Semitic writings affect and enrich the theory of writing”?

In this paper, Semitic writings will be exemplified by their latest-born system, that of Arabic. We present a cognitive view of the phonographic vs. graphic-to-meaning relations. Both aspects are concerned with the production vs. recognition and the writing vs. reading processes, considered in relation to the emergence of Semitic writing systems. This approach will allow us to analyse and describe the emergence and codification of the Arabic writing system, in a renewed synthesis.

## 2. A Traditional View Calling for a Deeply Renewed Approach (Recall)

The fact that Semitic writings do not note short vowels in the body of words is an ancient and traditional question. It has been repeatedly asked in books on the history of writing, such as Marcel Cohens's, James Février's, I.J. Gelb's, and many others, as well as in the field of Semitic studies (G.R. Driver, D. Diringer...). It still remains repeated in many of today's books on writings. Ancient Greek and Latin writings have been described as *scripto continua*, since they did not separate words, and Semitic writings as *scriptio defectiva*, because they allegedly did not note vowels (except as diacritic signs, which were in fact introduced much later).

A first answer to that question—also repeatedly reproduced – mentions the weight of age-old tradition, and assumes that vowels were just not conceptualised<sup>1</sup>, which resulted in Semitic scripts being described,

---

1. In I.J. Gelb's view, conceptualising vowels is assumed to be a specific ability, which the Semitic peoples would have been unable to develop, due to the weight of

in a factually improper wording, as “consonantal writings” (I.J. Gelb 1963). The latter have been considered as a step in the mythic idea of the “history of writing”,<sup>2</sup> i.e., in the “development of the invention of writing” (M. Cohen), the ultimate result of which would have been the ancient Greek alphabet. Such a view can no longer be held, because the multiple developments of writings in various cultures are now better known, and because the now prevalent idea is that every writing system is related to the language and culture in which it emerges and develops (see, e.g., J. Lyons, 1968, F. R. Harris, 1986, 1993, N. Catach (ed.), 1988, S.R. Fischer, 2001, F. Coulmas 2002, and, regarding Semitic writings with special reference to Arabic, J. Dichy, 1990, 2019).

### 3. Why This View Does Not Hold

Factually, it is essential to remember that Semitic writings had noted long vowels at the end, then in the body of words from the 13th century B.C. onwards (these are known as the *matres lectionis* of the Bible). This fact makes it difficult to go on describing these writings as “consonantal”.

Epistemologically, this view pertains to the illusive idea of the “development of writing”, which goes back to the 18th century (Warburton, 1744). These views can be opposed to the analytic approach of Condillac (1746; 1775) (Dichy, 2017). We cut this discussion short here, in order to base our hypotheses on precise facts from the Semitic and Arabic writing systems.

### 4. Another Still Current Inadequate Hypothesis

Another hypothesis, initially brought forward by James Février and Marcel Cohen, has been widely taken up by Arab linguists. It suggests that Semitic morphological patterns make up for the lack of vowels.

The main trouble with this hypothesis is that it only covers a percentage of words (for Arabic, see Dichy, 1992). Although all verbs and basic deverbal forms (such as the infinitive, *masdar*, the active and passive participles, *ism al-fā'il wa-l-maf'ūl*, the analogous adjective, *sifa mushabbaba*, etc.) are based on morphological patterns, the choice of the pattern found in a given sentence does not allow in many cases the determination of the vowels, because the same sequence of letters can be shared

---

tradition. We will see below that this was a factually inadequate observation. One needs to recall these points, because Gelb's synthesis on the development of writing systems still has an impact on a number of linguists.

2. The singular form of “writing” and “invention” is significant.

by various patterns, e.g., the sequence  $\text{يَعْلَم}$  /yʿlm/ can be instantiated as  $\text{يَعْلَم}$  *yaʿlamu*, ‘he knows’ or as  $\text{يُعْلِم}$  *yuʿlimu*, ‘he informs’, or  $\text{يُعَلِّم}$  *yuʿallimu*, ‘he teaches’, to which three other ‘passive-form’ sequences can be added.

In addition, a high number of nouns is not based on a predictable morphological pattern.

## 5. Remaining Questions: Why and How Are Short Vowels Not Written in Standard Script?

One must note that what is commonly called “unvowelled writing” should be more correctly described as “non-diacriticised” script. *Secondary diacritics*<sup>3</sup> essentially include the following signs:

- short vowels,
- consonants doubling (*shadda*),
- case-ending diacritics for both determined and undetermined nouns
- mute consonant symbol (*sukûn*)...

The key to understanding this issue is the non-symmetrical relation between the reading and the phono-graphic processes. Such processes are related to the fact that writing is a *socio-cultural artefact*, stemming from the institution of norms in a given society. Historically, a language is submitted, after it first appears to a *grammatisation* process (R. Balibar 1985; Sylvain Auroux 1994). In the Arabian culture, such a process occurred in the first three to four centuries after the emergence of Islam, and especially in the first period of the Abbasid dynasty.

## 6. The Emergence and Development of a Writing System

The general concept is that of the *Analytics of writing* (Dichy, 2017), according to which the emergence of a writing system stems from the way in which a given culture analyses the structures of its own

- produces finite inventories of phono-graphic units and morpho-graphemic structures on the one hand,
- and projects them on the support of writing through a system of codified forms combining letters and word-forms on the other hand.

Let us consider both aspects.

---

3. *Primary diacritics* are dots added over or under letters of the same shape, to identify the grapheme referred to, such as  $\text{ب} - \text{ت} - \text{ث}$ , respectively *y-b-t-t-n*. These dots are written with the thick end of the calamus, which shows that they actually belong to the letter, while the secondary diacritics are drawn with the thin end (Dichy, 1990, 2019).

## 6.1. The Phono-Graphic Perspective, in Relation to the Original Emergence of the System

In the first Western and Southern Semitic writings, the basis of the script was the inventory of the letters, which was later described as an *alphabet*, i.e., a set of grapho-phonemes. The latter are identified through an *intuitive phonology*, which breaks down syllables into smaller units. One then needs to describe the way in which such fundamental units are determined. Two ‘intuitive’ criteria will be highlighted or recalled here, respectively, the phonographic convention and the relation between letters and root-consonants.

### 6.1.1. *The Phonographic Convention*

As I had extensively shown for Arabic, the phono-graphic convention which resulted in the inventory of letters—i.e., of graphemes noted in the body of words—is essentially rhythmical. The convention can be phrased in very short words as follows:

Write a letter for the initial of every syllable, adding a second letter if the syllable is long (i.e., of the CVC or CVV form, where C is for “consonant”, V is for “vowel” and VV for “long vowel”).

This convention is directly related to the syllabic structure of these languages, the fundamental syllables of which are:

- CV – CVC/CVV     *ma – man/maa.*

In addition to these basic syllables, contextually determined ones appear:

- CVCC – CVVC     *mart – baab.*

One must remember that all syllables in this family of languages begin with a consonant. One could describe the resulting writing as metric/rhythmic, example:

*Samar taaliba* سمر طالبة ‘Samar [is] a student’, where & is for the final feminine ending, and *A* in the tables below, for the letter *alif*, which notes the second half of the long consonant *aa*. Case-ending are omitted in this example, as in standard speech.

In Table 1, capital transcription letters stand for letters appearing in the body of words in Arabic writing.

The example of Table 1 features the ‘intuitive phonology’ that led to the inventory of letters.

TABLE 1. Rhythmic/metric structure of Arabic writing

| Sa                            | MaR   | TaA  | Li                            | Ba  | & |
|-------------------------------|---|--|-------------------------------|---|---|
| Cv                            | CvC   | CvV  | Cv                            | CvC   |   |
| Short syllable                | Long syllable<br>(ending with<br>a consonant) | Long syllable<br>(ending with<br>the 2nd half<br>of a long<br>vowel) | Short syllable                | Long syllable<br>(ending with<br>a consonant) |   |
| 1 metric/<br>rhythmic<br>unit | 2 metric/<br>rhythmic<br>units                | 2 metric/<br>rhythmic<br>units                                       | 1 metric/<br>rhythmic<br>unit | 2 metric/<br>rhythmic<br>units                |   |
| 1 letter                      | 2 letters                                     | 2 letters  | 1 letter                      | 2 letters                                     |   |
| S                             | MR  | TA   | L                             | B&  |   |

### 6.1.2. *The Relation Between Letters and Root-Consonants*

In addition to the rhythmic indication related to the phonology of Arabic mentioned above, one must note that Semitic roots are always consonantal, while Indo-European languages, including ancient Greek, feature vocalic syllables, which partly accounts for the fact that, upon adopting the Semitic alphabet, ancient Greek has added vowels to its basic inventory of letters<sup>4</sup>.

In Arabic writing, roots strictly remain consonantal, even when they include vocalic consonants *w* (و) or *y* (ي) although these letters are also used for long vowels. The case of *alif* (ا) which only notes the long vowel *â*, as in *bâb* (باب) ‘door’, or *lâ* (لا) ‘no’, is significant: *alif* (ا) is never included in a root (i.e., as a radical). The Arabic alphabet thus only includes consonants, the first letter, which is *alif* (أ), refers in fact to the glottal stop *hamza* (ء)<sup>5</sup>, according to the principle that the first letter of a name of the unit of the alphabet corresponds to the sound denoted by it (this is known as the principle of acrophony). In his comment on the name of the *alif*, Ibn Jinnî (10th/4th century) thus recalled that the name *jîm* (جيم) referred to the letter *j* and not, for instance to *m* albeit it includes the sound (Sirr Sinâ‘at al-‘i‘râb, سر صناعة الإعراب).

In the ‘intuitive morphology’ in consideration, the inventory of letters thus appears to be related to consonantal roots.

4. M. Cohen (1958) suggested that the existence of vocalic roots explained the adding of vowels by the ancient Greeks to their alphabet.

5. The name *hamza* does not belong to the traditional alphabet. The corresponding written symbol (ء) has been added later.

## 6.2. The Morpho-Graphic Recognition or Reading Perspective

Let us now move to a presentation of the complementary aspect, related to the word-form and the reading perspective.

### 6.2.1. *The Word-Form Structure*

The complex structure of the word-form in Arabic can be represented as follows: a lexical unit appears at the centre of the word-form, the structure of which includes two sets of grammatical formants positioned right and left of a lexical knot (or stem). These formants appear in two layers. The first one is necessary to the morphological structure of the word-form, and results in what can be described as a *minimal word-form* (D. Cohen, 1961), as can be seen in Table 2.

TABLE 2. Minimal Arabic word-form structure

| Layer 1: Minimal word-form            |        |              |               |
|---------------------------------------|--------|--------------|---------------|
| Word formants                         | PREFIX | LEXICAL STEM | SUFFIX        |
| Verb, vowels included                 | Ya     | NZIL         | uWNa          |
| Translation or grammatical indication | 'they' | 'go down'    | masc., plural |
| Written form in standard writing      | Y      | NZL          | WN            |

Arabic word-forms can, in addition to suffixes and prefixes, comprehend proclitic and enclitic formants resulting in what can be called a *maximal word-form* (D. Cohen *op. cit.*), as can be seen in Table 3.

TABLE 3. Maximal Arabic word-form structure

| Layer 2: Maximal word-form            |                 |        |              |               |                                  |
|---------------------------------------|-----------------|--------|--------------|---------------|----------------------------------|
| Word formants                         | PROCLITICS      | PREFIX | LEXICAL STEM | SUFFIXES      | ENCLITICS                        |
| Word-form, vowels included            | Wa-Li           | Ta     | SKUN         | uW            | HaA                              |
| Translation or grammatical indication | 'And – so that' | 'You'  | 'inhabit'    | masc., plural | 'it' (in Arabic, fem., singular) |
| Written form in standard writing      | FL              | T      | SKN          | W             | HA                               |

Suffixes and prefixes on the one side, and proclitics and enclitics on the other, strictly belong to a *word formant grammar*. The inventory of

formants included in the fields positioned right and left of the lexical stem is, of course, limited<sup>6</sup>. This results within the reading process in a very structured word-form recognition set of operations.

### 6.2.2. *The Structure of the Lexical Stem and of Word-Forms Recognition*

Better to understand the above recognition process, one must remember that the lexical stem of the word-form can be analysed into ROOT and PATTERN in 100% of verbal forms, and a high percentage of nouns (Dichy 1990; 1992).

The word-form grammar is composed of rules relating the suffix, prefix proclitic and enclitic formants. It also includes rules linking these formants to the lexical stem<sup>7</sup>. It is to be noted that the enclitic formants include complement pronouns, cliticised to verbs, but also to nouns (the construct-state of which includes a pronoun in the second position).

Word-form recognition therefore involves the process of identifying the grammatical formants situated right and left of the stem. Every lexical stem is associated with *grammatical specifiers* combining the formants which can come right and left of it. For instance, transitive verbs accept complement pronouns as enclitics; some nouns accept the relative noun-adjective suffix *-iyy* (يَ) etc. These grammatical specifiers belong to the lexical features of the stem and are subsequently included in the word-form recognition process<sup>8</sup>.

### 6.2.3. *The Graphic Structure of the Word-Form and the Reading Process*

The final form of letters, which occurs in a small number of them in Aramaic and Hebrew, has been generalized in the writing system of Arabic. Word-Forms, which were usually separated by dots in ancient Semitic writings, are consequently recognizable in Arabic, where their borders are rendered visible by the final form of letters. These are systematically followed by a space in modern scripts and are often recognizable in ancient manuscripts.

---

6. A summary of the word-form grammar has been presented in (Dichy 1997), the complete rules of which have been developed in my 1990 work (chap. 10).

7. A limited number of stems are grammatical, such as *wa-inna-bumâ* (وإنهما), "and-that-two of them". A specific word-form grammar has been devised for them.

8. In the first half of the 1990's, 129,000 Arabic lexical stems have been associated with their word-form grammatical specifiers (after Dichy 1990) in the DIINAR (DIctionnaire INformatisé de l'ARabe) project, in a collaboration between Lyon and IRSIT (Institut de Recherche en Sciences Informatiques et des Télécommunications), a high-level Tunisian centre (Dichy, Braham, Ghazali & Hassoun 2002; Dichy & Hassoun 2005).



In addition, Arabic script, following Syriac writing, organizes words along a thick line, which is interrupted by the final form of letters, further identifying<sup>9</sup> word-form boundaries.

Example, 'Imru'u l-Qays's verse:

|   |   |
|---|---|
| فأشئت من شعرهن اصطفت                                    | تخيرني الجن أشعارها                           |
| <i>tukbayyirunī l-jinnu 'asb'ārābā</i>                  | <i>fa-mā shi'tu min shi'ribinna STafaytu</i>  |
| <i>The Djinns allow me choice between their rhymes,</i> | <i>Whichever verse I choose I may retain.</i> |

This word-form structure entails a 'contour' reading of words in Arabic (Grainger, Dichy et al., 2003) as well as in Hebrew (Frost, Forster & Deutsch, 1997, 2000). Words are subsequently recognized in a different process than the one we know in Latin character writings, where a word becomes recognizable after the second, third or fourth letter in most cases. Of course, Latin character writings combine contour and letter-by-letter recognition, as opposed to Semitic writings, which are fundamentally based on the contour recognition and the analytic processes of the word-form.

## 7. Is Unvowelled Writing a 'Robust' Abstraction With Regards the Reading Process?

Another hypothesis has been brought forward (Dichy, 2017).

J. Vendryes (1923) opposed the idea that writing should be a mirror image of phonetic realisations on the basis of the fact that pronunciation varies, sometimes strongly, from one region to another within the same language. He considered that writing needed to reflect a type of phonological abstraction shared by speakers whose pronunciations varied. We have seen above the abstraction based on intuitive phonology that Semitic writings reflect.

Considering the level of variation of ancient West-Semitic languages (Phoenician, proto-Hebraic, Eblaite, etc.), it is highly probable that the realisation of many short vowels differed. An additional hypothesis could then be that the intuitive phonology underlying the writings of these Semitic languages resulted in "robust" scriptural systems, i.e., in writing systems featuring a level of abstraction allowing them to be shared by a variety of dialects or a-kin languages.

This hypothesis nevertheless encounters two general objections:

9. Four letter forms interrupt the line in the middle of the word. These are *dāl*, *wāw*, *alif*, *rā'* (د، و، ا، ر), to which *dhāl* (ذ) and *zāy* (ز) must be added. This question is related to the history of the writing system of Arabic (Dichy 1990).

1. It only concerns a part of the writing data, since the diacriticisation system of Arabic includes, as mentioned above, other symbols than that of short vowels.
2. It remains based on a mistaken view of the reading process, which should not be considered as a writing-to-sounds, but as a reading-to-meaning activity.

One must also add that this feature of the writings systems in consideration should not be considered teleologically, being a consequence of the processes presented above, and not a feature these writings could have been “devised for”.

## 8. A Summarized Answer to the Short Vowels or ‘Scriptio Defectiva’ Issue

The answer to the so called ‘*scriptio defectiva*’ issue, in other words to that of the short vowels in Semitic writing, based on the above short presentation of the structure of Arabic writing, includes the following complementary answers:

1. Due to the phonological structure of Semitic languages, according to which all syllables begin with a consonant, the phono-graphic inventory of letters is based on an intuitive metrical/rhythmical analysis of spoken utterances. This results in Arabic in the notation of long vowels in the body of words and the omission of short vowels,
2. The fact that all Semitic roots are consonantal can be considered as a complementary reinforcement of the above phono-graphic structure.
3. The word-form structure, which we have described above as entailing a contour recognition of word-forms, involves a recognition process that does not call on the mediation of sounds for reading, in addition to information related to the syntactic structure and the context.
4. The reading process in Arabic proves to allow the understanding of written texts at the same level of efficiency as one encounters, say, in English<sup>10</sup>. Native Arabic speakers do not consider the short vowel issue as an impediment for either writing or reading Arabic texts. They often refuse a systematic notation of the short vowels and other secondary diacritics, except in religious or ancient literary and poetic texts.

---

10. On the other hand, correct reading aloud of Arabic texts is more difficult than the actual reading-for-meaning process. In teaching Arabic both as a national and foreign language, education programs as well as teachers still most often mix up reading-for-meaning and reading aloud, which may result in inefficient teaching of the written language.

5. The idea that unvowelled writing systems could be considered as 'robust' with regards to dialectical variation, does not present us with an explicative hypothesis, albeit it may partially be retained.

## 9. Conclusive Remarks

Taking up the question put forward in the first lines of this work, about whether the analysis of Semitic writings exemplified by Arabic could affect and enrich the theory of writing, one can observe that:

- (1) The writing system of Arabic features strong reading-for-meaning processes, essentially based on the structure of word-forms, the centre of which is—except for grammatical words—a lexical unit.
- (2) These lexical units—or stems—are associated with morpho-lexical specifiers that relate them—through a *word-form grammar*—to the other formants encompassed in the word-form.
- (3) Word-form boundaries are rendered visible by the final form of letters that interrupt the line along which letters are drawn within the word-form.
- (4) Semitic writings can by no means be reduced to "consonantal" scripts devised by peoples that did not come to the level of conceptualization reached by the ancient Greeks (who added vowels to their alphabets). These writings included, from the 13th cent. B.C. onwards, long vowels.
- (5) The inventory of letters included in Semitic alphabets stemmed from an intuitive phonology that can be described as rhythmical/metrical-sensitive, due to the structure of syllables that always start with a consonant system, in addition to the fact that the roots of these languages are exclusively consonantal.

The features presented in this paragraph and the previous one illustrate the way in which the concept of the *Analytics of writing*, which considers the analysis of spoken utterances through an 'intuitive phonology' leading to an inventory of letters (in the case of Semitic conventional graphic system). These analytics include the identification of lexical units and their projection on written realisations. In Semitic writings, which have always visually represented word-form boundaries, this results in a word-form recognition process, which we have recalled for Arabic.

## References

- Auroux, Sylvain (1994). *La Révolution technologique de la grammatisation*. Liège: Mardaga.

- Balibar, René (1985). *L'Institution du français. Essai sur le colinguisme des Carolingiens à la République*. Paris: PUF.
- Catach, N. (1988). "L'écriture en tant que plurisystème, ou théorie de L. Prime." In: *Pour une théorie de la langue écrite*. Ed. by N. Catach. Paris: Éditions du CNRS, pp. 243–259.
- Catach, Nina (1980). *L'orthographe française. Traité théorique et pratique*. With the collaboration of Gruaz, C. and Duprez, D. Paris: Nathan.
- Cohen, David (1961). "Essai d'une analyse automatique de l'arabe." In: *Études de linguistique sémitique et arabe*. Ed. by D. Cohen. Reprod. in D. Cohen, *Études de linguistique sémitique et arabe*. The Hague/Paris: Mouton, pp. 49–78.
- Cohen, Marcel (1958). *La grande invention de l'écriture et son évolution*. Vol. 3. Paris: Imprimerie nationale and Klincksieck.
- Condillac, E. Bonnot de (1775). *Grammaire*. Ed. by G. Le Roy. Paris: Presses Universitaires de France.
- (1798). *Essai sur l'origine des connaissances humaines*. Ed. by G. Le Roy. rev. ed. Vol. 1. 1746, rev. ed. 1798. Paris: Presses Universitaires de France.
- Coulmas, Florian (2002). *Writing systems. An introduction to their linguistic analysis*. Cambridge University Press.
- Dichy, Joseph (1990). "L'écriture dans la représentation de la langue: la lettre et le mot en arabe." Doctorat d'État. 2 vol. PhD thesis. Lyon: Université de Lyon, to appear in *Grapholinguistics and Its Applications*, Brest: Fluxus Editions.
- (Dec. 1992). "Pourquoi l'écriture arabe ne note pas les voyelles brèves." In: *Les systèmes d'écriture, Liaisons HESO 21-22*, pp. 31–53.
- (1997). "Pour une lexicomatique de l'arabe: l'unité lexicale simple et l'inventaire fini des spécificateurs du domaine du mot." In: *Meta 42.2*, pp. 291–306.
- (2017). "The Analytics of Writing, Exemplified by Arabic, the Youngest of the Semitic Scripts." In: *Approaches to the History and Dialectology of Arabic, in Honour of Pierre Larcher*. Ed. by M. Sartori, M.E. Giolofo, and Ph. Cassuto. Leiden/Boston: E.J. Brill, pp. 29–56.
- (2019). "On the Writing system of Arabic: the Semiographic Principle as reflected in Naskhi letter shapes." In: *Graphematics in the 21st Century. Proceedings of the /guafematik/ 2018 Conference*. Ed. by Yannis Haralambous. Vol. 1. Grapholinguistics Series.
- Dichy, Joseph and Mohamed Hassoun (Apr. 2005). "The DIINAR.1-«معالي» Arabic Lexical Resource, an outline of contents and methodology." In: *The ELRA Newsletter 10.2*, pp. 5–10.
- Dichy, Joseph et al. (2002). "La base de connaissances linguistiques DIINAR.1." In: *Colloque international sur le traitement automatique de l'arabe*. La Manouba-Tunis, pp. 45–56.
- Diringer, David (1968). *The Alphabet. A key to the History of Mankind*. 3rd revised ed. New York: Funk & Wagnalls.

- Driver, Godfrey R. (1976). *Semitic Writing: From Pictograph to Alphabet*. 3rd revised ed. Oxford: O.U.P.
- Février, James (1959). *Histoire de l'écriture*. 2nd ed. Paris: Payot.
- Fisher, Steven R. (2001). *A History of Writing*. London: Reaktion Books.
- Frost, R., A. Deutsch, and K. Forster (2000). "Decomposing morphologically complex words in a non linear morphology." In: *Journal of Experimental Psychology: Learning, Memory and Cognition* 26, pp. 751–65.
- Frost, R., K. Forster, and A. Deutsch (1997). "What can we learn from the morphology of Hebrew? A masked priming investigation of morphological representation." In: *Journal of Experimental Psychology: Learning, Memory and Cognition* 23, pp. 829–856.
- Gelb, Ignace J. (1963). *A Study of Writing*. London and Chicago: University of Chicago Press.
- Gleason, H. A. (1961). *An Introduction to Descriptive Linguistics*. New York: Holt, Rinehart & Winston.
- Grainger, Jonathan et al. (2003). "Approche expérimentale de la reconnaissance du mot écrit en arabe." In: *Dynamiques de l'écriture: approches pluridisciplinaires*. Ed. by Jean-Pierre Jaffré. 22, pp. 77–86.
- al-Hagib, Ibn (1975). "الشافية في علمي التصريف والخط [Al-Shāfiya in the Sciences of Inflection and Calligraphy]." In: *الأستراباذي، شرح الشافية [al-Astarābādī, Explanation of al-Shāfiya]*. Ed. by Nur al-Hasan et al. Vol. 4. Beirut: دار الكتب العلمية [House of books of science].
- Harris, Roy (1986). *The Origin of writing*. London: Duckworth.
- (1993). *La sémiologie de l'écriture*. Paris: CNRS Editions.
- Healey, John F. and G. Rex Smith (2009 and 2012). *A brief introduction to the Arabic Alphabet: its origins and various forms*. London, San Francisco, Beirut: Saqi.
- Ibn Jinni, Abu l-Fath Utman (1985). *سر صناعة الإعراب [The Secret of the Craft of Inflection]*. Ed. by Hassan Hindawi. Vol. 2. Damascus: دار القلم [House of the Qalam].
- Jaffré, Jean-Pierre (1988). "Graphèmes et idéographie. Approche psycholinguistique de la notion de graphème." In: *Pour une théorie de la langue écrite*. Ed. by N. Catach. Paris: Éditions du CNRS, pp. 93–102.
- Lyons, John (1968). *General Linguistics*. § 2.2.6. Cambridge University Press.
- Olson, David R. (1994). *The World on Paper: The conceptual and cognitive implications of reading and writing*. Cambridge University Press.
- Osborn, J.R. (2017). *Letters of Light. Arabic Script in Calligraphy, Print, and Digital Design*. Cambridge (Massachusetts) and London: Harvard University Press.
- Sampson, Geoffrey (1985). *Writing Systems. A linguistic introduction*. Stanford University Press.

- Vendryès, Joseph (1923/68). *Le langage, Introduction linguistique à l'histoire*. rééd. Paris: Albin Michel.
- Vygotsky, Lev S. (1934). *Thought and Language*. Trans. by F. Sève. Quoted here after the French ed. *Pensée et langage*. Transl. by F. Sève. Paris: Messidor/Editions sociales, 1985. Cambridge, Mass.: MIT Press.
- (1935). "Le problème de l'enseignement et du développement mental à l'âge scolaire." In: *Vygotsky aujourd'hui*. Ed. by B. Schneuwly and J.P. Bronckart. Neuchâtel/Paris: Delachaux/Niestlé, pp. 95–117.
- Warburton (1744). *The Divine Legation of Moses Demonstrated*. Quoted after the incomplete French transl. by L. de Malpeines, *Essai sur les hiéroglyphes des Égyptiens*. Paris: Guérin.

# Reasons for Re-Paragraphing in the Translation Process

## An Ongoing Project

Dana Awad

*Abstract.* In this paper, we present ongoing research to establish an understanding of re-paragraphing in the translation process. By re-paragraphing, we mean changing paragraph structure or paragraph size and number in the target text; in other words, the decision to translate one paragraph into two paragraphs or vice-versa. We look into possible reasons from a syntactic point of view, and we suggest elements that would help set standards for translators to ensure a loyal transmission of text coherence.

### 1. Introduction

A paragraph is a universal concept in all languages; it is a textual unit with a topic and represents an idea in a text. In that sense, starting a new paragraph connotes a new idea. In this perspective, we aim to understand the reasons behind changes in paragraph number or paragraph division in a target text. As translation is the transfer of meaning from a source language to a target language, what comes to mind is transferring the meaning of words depending on the context, translating sentences, and adapting sentence structure in the process if deemed grammatically necessary due to syntactic and grammatical differences between languages. However, how does re-paragraphing contribute to transferring meaning when it comes to changing paragraph divisions? In a text, is changing paragraph structure or paragraph number considered necessary to ensure an accurate transfer of the meaning of a text, of its logical organization of ideas? We aim to find out a paragraph's role in the transfer of meaning and to explain if re-paragraphing is sometimes necessary for a successful translation. Even though re-paragraphing does not occur as much as changing the syntactic structure of a sentence, which happens for grammatical reasons, this shift is intended. It is considered a necessity to achieve 'naturalness' in the target text.

---

Dana Awad  0000-0001-9522-4549

Prince Sultan University, Applied Linguistics Research Lab (ALLAB)

E-mail: DanaAWAD@hotmail.fr

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2022. Proceedings*  
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 9.  
Fluxus Editions, Brest, 2024, pp. 391–398. <https://doi.org/10.36824/2022-graf-awad>  
ISBN: 978-2-487055-04-9, e-ISBN: 978-2-487055-05-6

In this paper, we will describe the concept of a paragraph and the different stages of paragraph analysis, and we will attempt to explain the reasons behind re-paragraphing using examples of re-paragraphing in Arabic translations. At this point of the research project, our objective is to have a general understanding of the reasons behind re-paragraphing regardless of text typology related to stylistics.

## 2. The Notion of a Paragraph From a Translation Perspective

The common definition of a paragraph is a section of a written text covering a specific theme. Each paragraph should have an introduction, a development, and a conclusion of its own, marked by a line space and an indentation. This is the universal definition of a paragraph as found in dictionaries, which can be summarized as such:

A distinct section of a piece of writing usually dealing with a single theme and indicated by a new line, indentation, or numbering.  
(Oxford English dictionary)

Some other definitions (Merriam Webster) add more precise details, saying that it is formed of a group of sentences or a single sentence that “forms a unit” and that has an introduction, development, or conclusion either directly, as in academic texts, or indirectly as in literary texts. The definition of a paragraph in Arabic is not different than the former definitions. In Arabic, a paragraph is defined as

A part of discourse or a written idea that covers a specific theme of a topic in a book or an article.  
(al-Maany dictionary)

جزء من كلام أو فكرة مكتوبة تتناول نقطة معينة من الموضوع كفقرة من كتاب أو مقال.

We find the former summary of definitions of a paragraph does not succeed in defining the concept of a paragraph from a linguistic point of view to differentiate it from the concept of written discourse. The only distinction of a modern paragraph is the alinea and indentation for some paragraphs. Even though the word (or term) paragraph globally has the same meaning, the concept of a paragraph remains too general since it implies too many factors for it to be as global as its definition implies. A paragraph is a linguistic entity in a discourse. It is accompanied by an alinea or indentation, which makes a paragraph an essential factor in determining a text layout. It is related to cognitive sciences for its role in clarifying the logic between ideas in a text and creating a ‘smooth transition’ between ideas or events in a given text, connecting or separating ideas from each other. Since it does not necessarily have linguistic markers, the segmentation of a text into paragraphs depends on the author and what seems natural in a language, which is also a vague notion.



From a translation perspective, the division of paragraphs is related to text coherence, which makes the meaning of a paragraph from a psycholinguistic perspective the most relatable to the translation process and the translator's decision to change paragraph structure or paragraph division. Coherence is somehow a vague concept because, unlike cohesion, no linguistic markers define coherence in a text. The first step when translating a text is to read it in its entirety to understand how the author logically relates ideas and then transfer this logic into a target language. Therefore, coherence transfer is an essential part of the translation process. The understanding and, as a result, the transfer of text coherence depends on the translator's understanding of the text and of the topic in general (Le, 2004)

This, however, does not exclude the importance of syntactic analysis, especially cohesive devices used to start a paragraph, which are also indicators of coherence in some languages, such as Arabic.

To have a structured analysis of a paragraph, we will consider it, at this point of our project, as a larger version of a sentence and, therefore, apply syntactic analysis of sentences in paragraph analysis of a written discourse; we will then try to define elements that would help set standards for re-paragraphing in translation.

### 3. Syntactic Analysis of a Paragraph: The Application of Transformational Grammar and Systemic Functional Grammar in the Translation of a Paragraph

In this syntactic analysis, we would compare the function of the source text and the source culture with the functions of the target text in the target culture. To achieve this, it is important to analyze a paragraph at the micro level (sentence order in a paragraph, intersentential relationships, and their involvement in the aesthetic form) and at the macro level (paragraph divisions and linguistic elements involved in marking the beginning of a paragraph).

At the micro level, analysis of connectors is essential to understand the relationship between sentences, such as coordination, subordination, or contrast. This micro-analysis of a paragraph is important in translating between Arabic and English because both languages have different norms in paragraph construction. In Arabic, building a paragraph with one long sentence with connectors is common. When translating such paragraphs into English, translators analyze connectors' role in building a paragraph with shorter sentences while maintaining intersentential relationships. See for example, Fig. 1.

In the example of Fig. 1, there is a two-sentence paragraph in English: the first is the statement of an example to support the following

And that is how, in worrying silence, the murder of the Paris schoolteacher Samuel Paty was used as a pretext for disbanding the Collective Against Islamophobia in France. **It's as if**, day after day, far from extending the limits of freedom, the explosion in communication is creating disciplinarian societies that force us to shuttle back and forth between our places of confinement.

على هذا النحو، جاء اغتيال صامويل باتي ليتخذ، في صمت مُريب، كعِلَّةٍ لحلّ الجمعية المناهضة لمعاداة الإسلام في فرنسا، كما لو أنّ طفرة وسائل الاتصال بدلا من التوسيع من نطاق الحريّات، تعتمد يوما بعد يوم إلى إرساء مجتمعات نظاميّة تحتم علينا مداومة التّقلُّل من وإلى مُحشّداتنا.

FIGURE 1

sentence, which is an argument (the sentence moves from a specific incidence to a general statement). The cohesion between both sentences is established with *It's as if*. In the Arabic equivalence, both sentences were translated into one sentence using the cohesive device <i>kamā</i>, used to coordinate two complete sentences to refer to the similarity between them (the literal translation is *as*). This change in the micro-construction of a paragraph is essential so that the reader can understand the author's intended coherence.

This structure of one-sentence paragraphs is common in Arabic, and transformational grammar (Chomsky, 1957) is a possible solution for translating one-sentence paragraphs. Since the comma, along with connectors, are often used as indirect sentence boundaries in a long Arabic sentence, they can give the translator hints on how to re-construct the meaning naturally, creating an appropriate number of sentences that would make the text readable in English while maintaining the source text's cohesion.

Another way to analyze a paragraph at the micro level is by applying systemic functional grammar (Halliday, 2014), especially theme-rheme organization.<sup>1</sup> In that sense, a paragraph is analyzed by theme-rheme sequence. In a paragraph, a theme is the topic sentence, and the theme is the supporting sentence. This type of analysis is internalized in the translators' text and paragraph analysis. Still, the theme-rheme organization of a paragraph changes when it is translated into two paragraphs, thus creating two topic sentences. Consider the example of Fig. 2, where the translation in English and Arabic have the same theme-rheme organization in two paragraphs but, in the French translation, both paragraphs were translated into one.

The Arabic and English versions have two topic sentences, while in the French version, both paragraphs were translated into one, the sec-

1. Researchers such as Fareh (1988) and Aziz (1988) applied theme-rheme organization in the broader sense of paragraph analysis.

Cependant, quand la droite américaine s'en indigne, on est presque tenté de lui répliquer: n'est-ce pas vous et vos penseurs de Chicago qui avez installé l'idée que la puissance publique ne devait brider ni le pouvoir des entreprises, ni la fortune de leurs propriétaires, légitimés selon vous par le libre choix des consommateurs? Eh bien, ce «populisme de marché», vous en devenez aujourd'hui les victimes. **Le premier amendement de la Constitution américaine protège la libre expression contre une censure de l'État fédéral et des gouvernements locaux, mais pas contre celle des entreprises privées en situation de monopole. Leur «expression», c'est votre silence. Vae victis, en somme, et tout le pouvoir aux Gafam (2) lorsqu'ils vous font taire!**

When the US right expresses outrage, however, one is tempted to reply: wasn't it you and your Chicago ideologues who established the idea that government should not limit the power of business enterprises or the wealth of their owners, which (according to you) were legitimized by consumer choice? Well, now you're the latest victims of this 'market populism.'

**The First Amendment to the US Constitution protects freedom of expression against censorship by federal or local government, but not against that of private enterprises operating a monopoly. Their 'expression' has become your silence. Woe to the vanquished, and all power to the GAFAM (2) when they shut you up!**

لكن وعندما يعبر اليمين الأمريكي عن سخطه على هذا الفعل، فإننا نبدو أميل إلى أن نرد عليه بقولنا: أَلَسْتَ أَنْتَ، بمعية مفكريك في شيكاغو، مَنْ وضع فكرة أن السلطة العمومية لا يمكنها أن تحد من سلطة المؤسسات الخاصة ولا من ثروة أصحابها، وهو الأمر الذي شرعتم له باسم حرية اختيار المستهلكين؟ حسنا، هاهي «شعبوية السوق» وها أنتم اليوم أصبحتم من ضمن ضحاياها.

يحمي التعديل الأول لدستور الولايات المتحدة الأمريكية حرية التعبير من رقابة الدولة الفدرالية والحكومات المحلية، لكنه لا يحميها من رقابة الشركات الاحتكارية الخاصة. إن «التعبير» عنها (تعبيرهم عن الرأي) يعني سكوتك. خلاصة الأمر: «لا عزاء للخاسرين»، ولغافام (2) مُطلق السلطة في منَعك من الكلام!

FIGURE 2

ond topic sentence being transferred into a supporting sentence. Even though this division does not happen often in the translation process, the possible reason would be cultural, the French reader would relate both paragraphs as part of one theme. Another example would be the following for a division of the Arabic paragraph into two, as in Fig. 3.

Figure 3 would be another example of translating one topic sentence into two topic sentences for socio-cultural reasons. The translator assumes that mentioning opponents of the war in Iraq and Afghanistan is important to the Arabic reader and, therefore, considers that the reader would find it more coherent if this sentence becomes the theme of a separate paragraph.

It becomes a continual and ever stricter state of emergency. Nothing is easier than identifying a target of hatred, shunned by all, and then continually extending the limits of censure and prohibition. **Opponents of the wars in Afghanistan and Iraq were labeled as Al-Qaida sympathizers; critics of Israeli policy as antisemites; and those who feel exhausted by the academic preachifying imported from the US as Trumpists or racists.** In such cases, we no longer seek to contradict our adversaries but to shut them up.

إنها لن تكون إلا حالة استثنائية تدوم وتشدّد. ذلك أنّه ليس هناك ما هو أيسرُ من تحديد هدف مكروه لا أحد يرغب في الظهور في صورة الشريك له، ثم العمل الدائب على توسيع محيط الرقابة والمحرمات. لقد تمّ وضّم المناهضين لحروب العراق وأفغانستان على أنّهم محامو تنظيم القاعدة—مناصرون لتنظيم القاعدة، كما نُعت منتقدو سياسة إسرائيل بأعداء السامية، أمّا من كانت الخطب الجامعية الوعظية المستوردة من أمريكا تثقلهم فيتهمون بكونهم من أنصار ترامب أو من العنصريين. لم يعد الأمر في مثل هذه الحالات متعلّقاً بمناقضة رأي الخصوم بل بإسكاتهم.

FIGURE 3

#### 4. How to Set Standards for Re-Paragraphing for Translation Purposes

To set standards for paragraph construction in a target text, we should define the linguistic elements involved in paragraph construction and their correspondence in different languages. If we consider the eight universals of discourse identified by Nida and Taber (in Steele, 1992, p.44), we consider the following essential to study cohesion and coherence in a paragraph:

- The marking of the beginning of discourse, which can be paragraph openings that are sometimes added as cohesive devices to mark the relationship with the previous paragraph).
- Temporal and spatial relations between events and objects.
- The identification of participants (theme in a topic sentence).
- The marking of logical relations between events (connectors at the micro level, punctuation marks).
- Highlighting emphasis.

From a linguistic perspective, we can use the aforementioned universals of discourse in the study of paragraph cohesion, which is the linguistic phenomenon concerned with the logic of a paragraph using explicit linguistic elements (Takagaki, 2008, p. 213). This study of paragraph cohesion can be at the micro level by analyzing connectors that explicitly show the logic between sentences, and at the macro level by analyzing cohesive devices that are sometimes used to highlight the logical arrangement of paragraphs in a text (paragraph openings). At the mi-

cro level, connectors that explicitly show the logic between sentences, and at the macro, cohesive devices that show the logical arrangement of paragraphs in a text (paragraph openings).

From a sociocultural perspective, the study of extra-linguistic elements that implicitly create coherence (Ibid.), which is the arrangement of ideas in an acceptable way according to the type of text and in a way that connotes a continuity of the thought process in a text can be achieved by identifying temporal or spatial relations between different elements and through the identification of participants.

Since textual organization does not have clear linguistic rules, paragraphs can be translated using the same structure of the source text without making linguistic mistakes. However, the final result would be a text that looks strange or unnatural because of the lack of the logical element. Translators, usually translating into their native language, are “subconsciously” aware of the syntactic and sociocultural elements that will make the text sound natural in the target language, even though there are no clear rules, and make changes in paragraph structure accordingly. This internalized knowledge comes from the knowledge of similarities and differences in marking the discourse universals present in paragraph construction. For example, Arabic has more paragraph openings than English, some of them untranslatable, such as *wāw*<sup>2</sup> at the beginning of an Arabic paragraph. Therefore, a translator might add a paragraph opening in Arabic that transfers the original inter-paragraph relation or might not translate an Arabic paragraph opening into English. Temporal and spatial relations and other logical relations between objects and events are respected in translation to stay loyal to the author’s intention. As for identifying participants and highlighting emphasis, we saw in previous examples that the translator might choose to get involved in text coherence for socio-cultural reasons so that the ideas’ connections are more logical. This involvement results in changing paragraph division and is target-reader oriented.

## 5. Conclusion

The study of a paragraph in general and a contrastive study of a paragraph for translation purposes, in particular, is complicated for many reasons. The main reason is the multidisciplinary of paragraph analysis, which includes linguistic, logical, and visual aspects of thematic representation. Another reason lies in the translator’s role as a loyal and ‘detached’ transmitter of information; the size of the original text has to be respected. Therefore, unlike the understandable and acceptable changes in paragraphs at the micro level for syntactic reasons, changes

---

2. *Wāw* is originally a coordinator that means (and), but that is used to create paragraph cohesion.

in the number of paragraphs must be for logical and cognitive (socio-cultural) reasons.

A change in the number of paragraphs is a deliberate action in the translation process, not only to add a logically natural sequence of ideas (part of the transmission of the author's logical plan in a target language) but also to create emphasis where the translator feels needed, usually for socio-cultural reasons (part of the translator's involvement in text creation). This paper was a presentation of an ongoing research project in which we hope to achieve a detailed contrastive analysis of paragraph structure to explain and set standards for the act of re-paragraphing in translation.

## References

- Abdullah, A. and Z. al-Obaida (2017). "Text concept in the Arab heritage: towards the integration of naqli and aqli approach." In: *Journal of Islamic Social Sciences and Humanities* 10, pp. 113–126.
- Adam, Jean-Michel (2018). *Le paragraphe: entre phrases et texte*. Paris: Armand Colin.
- al-Amīn, B. (2018). "The grammar of the text in the Arabic linguistic heritage: Synthesis theory model." In: *Revue La phonétique* 20.3, pp. 635–647.
- Aziz, Y. (1988). "Theme-rheme organization and paragraph structure in standard Arabic." In: *Word* 39.2, pp. 117–128.
- Bessonnat, D. (1988). "Le découpage en paragraphes et ses fonctions." In: *Pratiques: linguistique, littérature, didactique* 57, pp. 81–105.
- Chomsky, N. (1957). *Syntactic structures*. Mouton.
- Fareh, I.F. (1988). "Paragraph structure in Arabic and English expository discourse." PhD thesis. Department of Linguistics, the University of Kansas.
- Halliday, M.A.K. and C. Matthiessen (2014). *Halliday's introduction to functional grammar*. 4th ed. Routledge.
- Halloun, M. (1987). "فواتح الفقرات في نصوص طه حسين" [Paragraph openings in Taha Hussein's texts]. In: *Bethlehem University Journal* 6, pp. 73–87.
- Le, E. (2004). "The role of paragraphs in the construction of coherence text linguistics and translation studies." In: *International Review of Applied Linguistics in Language Teaching* 42.3, pp. 259–275.
- Steele, M.H. (1992). "Paragraph structure and translation: the theory and practice of paragraph and other high-level structures in English and Russian narrative and the effect of the translation process upon these structures." PhD thesis. University of Glasgow.
- Takagaki, Y. (2008). "Les plans d'organisation textuelle en français et en japonais: de la rhétorique contrastive à la linguistique textuelle." PhD thesis. Université de Rouen.

# Index

- abjad, 29, 316, 317, 323, 325, 328, 332, 334, 433, 643
- abugida, 62, 77, 327, 334, 419, 422–424, 429, 432, 433, 448, 702–705, 707, 723
- Afrikaans, 316
- akshara, 47, 52–57, 62, 63, 80, 419–421, 423–435
- allography, 16, 56, 321, 692
- alphasyllabary, 62, 419, 422, 423, 429, 432, 433
- American Sign Language, 47, 57, 284, 285
- Annotorious, 289
- Arabic, 68, 69, 73, 113, 201, 202, 239, 245, 315–337, 339–341, 343, 345–347, 349, 351–355, 357, 359, 361, 363, 365, 367, 369, 371, 373, 375, 378, 379, 381–387, 392–395, 397, 400, 421, 432, 433, 616, 644, 772
- Ardhanagari, 325
- ASCII, 268
- asemic writing, 111, 115, 116, 118, 119, 121, 122, 124, 126
- asemiosis, 111, 113–116, 119, 122, 124
- Assyriology, 280, 290
- ATF, 268, 269, 288
- Athabaskan, 59
- Aviva, 615, 617, 619
- Azerbaijani, 322, 331–333, 337, 339, 343–345, 355, 358, 360–362, 364
- Aztec, 75, 79, 81–83, 86, 87, 89, 91, 95, 194, 513, 515, 578
- Balochi, 322
- bamboo, 690, 691, 711, 713, 719, 721, 728
- Bangla, 55, 446, 448, 449, 457–459, 462–465
- bark beetle, 120, 131
- Barthes, Roland, 115, 117, 118, 125
- Bembo, Pietro, 626
- Ben-Yehuda, Eliezer, 612
- Bigelow, Charles, 635
- Bodoni, Giambattista, 630, 631
- Borges, Jorge Luis, 121, 122
- Boro, 400, 437–439, 444–446, 448, 450, 452–456, 459–462
- boustrophedon, 472, 473, 476, 478, 484
- Brahmic, 45, 47, 52, 54, 55, 57, 63, 64, 318, 325, 327, 334–336, 420, 433, 702, 704, 722
- Braille, 433
- brevigraph, 299, 309, 310
- brush stroke, 654
- Carrier, 47, 57, 59–63, 433
- Chaim, 611–613, 615–619, 621, 623

- Champollion, Jean-François, 193–195, 202, 213
- Cherokee, 451
- Chiang, Ted, 649, 650
- Chinese, 6, 45, 47, 48, 51, 63, 73, 78, 95, 112, 135, 136, 138–147, 149–153, 160–164, 166, 167, 169, 171, 172, 176–182, 184, 185, 194, 212, 266, 267, 271, 272, 274, 275, 279, 286, 316, 578, 645–649, 657, 671–677, 679, 681, 683, 685–687, 689–709, 714, 715, 717, 720–723, 725, 726, 739–751, 753, 755–757, 759, 761, 763–765, 767–769, 771, 777, 778
- circumfix, 47, 55, 64, 87
- Cohen, Marcel, 6, 377–379, 382, 383, 513
- Comic Sans, 105
- commercial sign, 135, 140, 153
- computer-mediated communication, 157, 158
- Coptic, 195, 196, 199, 200, 202, 205, 208–210, 212, 258, 562
- Courier New, 104
- cranberry morpheme, 741
- Cree, 59, 60, 64, 433, 673, 699
- critical theory, 439
- cuneiform, 201, 228–230, 265–283, 285–295, 297, 570
- cursivization, 316, 709
- Cyril, 240, 241, 246, 256–258
- Cyrillic, 29, 69, 237–239, 241, 243, 245, 247–251, 257, 330, 332, 452, 616, 772, 777
- da Vinci, Leonardo, 119
- Dari, 322
- de Sacy, Silvestre, 194–199, 201, 202, 209–212
- decipherment, 193–195, 197, 199, 206, 208, 210–214, 471, 550, 690
- derivational morphology, 53, 54, 59
- Derrida, Jacques, 2, 5–7, 10, 14, 29, 30, 34, 256, 700
- Devanagari, 52–55, 79, 80, 322, 325–327, 421, 424, 425, 428, 429, 438, 444–446, 453–456, 459
- digraphia, 67–69, 71, 73, 224, 225
- disfluency, 101–107, 109
- DjVu, 299, 307–309, 312
- Easter Island, 471, 473–475, 477, 479, 481, 483, 485, 487, 489, 491, 493, 495, 497, 499, 543, 548
- Eblaite, 385
- emblem, 75, 77, 79, 81–87, 89, 91–95, 97, 99, 419, 513, 530, 551
- Emojigeddon, 304
- Emojipedia, 165
- English, 4, 8, 15–17, 20, 22, 26, 27, 29, 30, 36, 56, 57, 59, 62, 72, 95, 114, 136, 144–148, 153, 160, 161, 172, 178, 181, 185, 199, 208–210, 225, 327, 334, 345–349, 352, 353, 386, 392–394, 397, 419, 420, 427, 440, 443, 453–455, 460, 462, 463, 489, 494, 506, 545, 549, 645, 646, 648–650, 657, 667, 672, 674, 689, 691, 701–704, 708, 740–742, 746, 772
- entaxis, 77, 79, 87, 91, 96
- entelechy, 717
- Facebook, 166, 255, 407, 410, 411, 446, 450, 452–456, 465
- Father Adrien Gabriel Morice, 59



- Février, James, 377–379  
 FIFA World Cup, 254  
 fleur-de-lys, 573, 574  
 Foucault, Michel, 698  
 French, 2, 4, 5, 7, 14, 15, 20, 26, 28, 29, 69, 71, 161, 172, 181, 194, 195, 199, 209, 221–223, 394, 395, 400, 445, 450, 564, 633, 672–674, 778  
 Friulian, 69  
 Galician, 69  
 Gallo-Italic, 69  
 Gallo-Romance, 28, 69  
 Gelb, I.J., 6, 378, 513  
 gender identity, 399  
 gender-neutral, 399, 401, 402  
 German, 2, 4, 11, 13–17, 19–22, 25, 26, 36, 37, 71, 144, 170–172, 176, 199, 221, 225, 248, 257, 265, 300, 400, 421, 449, 481, 491, 492, 548, 576, 643, 647, 657, 739, 778  
 Glagolitic, 237–245, 247–251, 253–263  
 glyph block, 47, 49–53, 64  
 gossip, 165, 167  
 grammaticography, 35  
 grammatogenesis, 726  
 grammatography, 4, 5  
 grammatology, 1, 2, 4–7, 10, 14, 30, 33–35, 256, 689, 694, 695, 697–700, 702, 704, 709, 717, 719, 723–726  
 graphematics, 5, 6, 14–17, 25, 37, 421, 440  
 graphemic cluster, 302  
 graphemics, 2, 8, 9, 11, 12, 14–17, 36, 37  
 graphiology, 12  
 graphology, 1, 5–15, 35, 692  
 graphonomics, 10  
 graphonomy, 1, 2, 6–10, 14, 35  
 Gutenberg, Johannes, 304, 625–628, 636, 646  
 Habermas, Jürgen, 449  
 hangul, 61, 77, 80, 81, 431, 433, 640, 651, 771, 772, 779  
 Hantology, 677  
 Hànzì, 45, 47–49, 51, 63, 64  
 HanziNet, 677  
 heterography, 425, 428, 429, 431  
 hieroglyph, 193–195, 199, 201–213, 231, 283, 284, 527, 543  
 Hjelmslev, Louis, 12, 76  
 Holmes, Kris, 635  
 Holmes, Sherlock, 641  
 ICANN, 320  
 illegibility, 112, 115, 116  
 Inka, 501–506, 508–512, 516, 519, 526–542, 544–550, 552–576, 578–583  
 Instagram, 173, 186, 411, 452–454  
 IPA, 60, 70, 465  
 Jangli, 324  
 Japanese, 48, 68, 73, 167, 172, 178, 179, 266, 267, 274, 316, 639, 641, 643, 645, 646, 653, 654, 657, 658, 663, 666, 668, 669, 709, 777  
 Jawi, 317, 322, 333, 334, 343, 345  
 JTF, 268, 269, 288, 289, 292  
 Unicode, 299, 300, 304, 306  
 juxtaposition, 85, 86, 550  
 Kangxi, 674, 675, 681, 686  
 Kashmiri, 322, 327, 328, 337, 339–341, 343–345, 355, 358, 359, 361–363, 421, 443  
 Kazuki, Miya, 645, 649, 651  
 kenogram, 431, 433  
 Khowar, 321, 328  
 Khudabadi, 325  
 kineticism, 113  
 Klingon, 639, 640, 644

- Kneser-Ney modeling, 340  
 Knesset, 409, 410  
 Kurdi, 328  
 Kurdish, 317, 322, 328, 329, 337,  
     339, 343–347, 349–355,  
     358, 359, 361–363, 544  
  
 Lahanda, 324  
 Leibniz, Gottfried Wilhelm, 197,  
     672, 699  
 Lewitt, Jan, 611  
 LGBTQ+, 400  
 Linear Elamite, 194  
 linked data, 266, 267, 269, 270,  
     272, 289, 292  
 liushu, 689, 691–693, 696–698,  
     701–703, 715–719, 721–  
     725  
 Lombard, 67–73  
  
 Mahajani, 325  
 Malagasy, 316  
 Malay, 52, 55, 67, 73, 317, 322, 333,  
     334, 337, 339, 340, 343–  
     345, 355, 358, 359, 361–  
     363, 426, 429, 443  
 Manipuri, 426, 437–439, 443,  
     446–449, 452, 453, 456–  
     466  
 Manuzio, Aldo, 628  
 maqaf, 401  
 Maya, 45, 47, 49–53, 63, 64, 286,  
     451, 494, 495, 548  
 Meitei Mayek, 440, 448, 449, 452,  
     457–459, 461–465  
 Methodius, 246, 256, 257  
 Michaux, Henri, 115, 117, 119, 130  
 Milanese, 70–73  
 Modern Standard Arabic, 319, 322  
 monographism, 67–69, 71, 73  
 More, Thomas, 639, 641–644, 648  
 morpheme, 23, 35, 46–49, 51, 55,  
     59, 75, 83, 84, 181, 182,  
     421, 540, 569, 740, 741,  
     743, 745, 747, 768  
  
 mouth action, 781–784, 786, 787,  
     793–798, 801, 803, 805,  
     806, 813  
 multigraphia, 68  
 multilingualism, 138, 150  
 musical notation, 702, 703, 707–  
     709  
  
 Nagari, 333  
 Naskh, 318, 319, 323–327, 329–  
     333  
 Nastaliq, 323–325, 327, 331, 332  
 nengoro, 661, 662  
 neographism, 692, 695, 696, 700,  
     719, 725  
 NFC, 318, 319, 335, 336, 340  
 nikud, 402, 405, 612  
  
 Ojibwe, 59  
 Old Church Slavonic, 237, 256,  
     257, 259  
 OntoLex, 265, 267, 269, 271, 273,  
     275, 277, 279, 281, 283,  
     285, 287, 289, 291, 293,  
     295, 297  
 ontology, 164, 265, 266, 269, 270,  
     273, 275, 279–287, 289,  
     290, 292, 293, 682, 696,  
     726  
 OpenType, 808  
 orthography, 1, 8, 9, 14, 16–18,  
     25, 35, 37, 67–72, 239,  
     251, 317, 319–322, 328–  
     330, 333, 354, 355, 421,  
     429, 437–444, 446, 451,  
     452, 456, 461, 464, 726,  
     727, 743, 747, 771, 773,  
     778  
  
 Pahawh Hmong, 77  
 Pallava, 333  
 Palmyrene, 197  
 Pashto, 322  
 Persian, 198, 201, 315, 317, 322,  
     323, 325, 330, 332, 333,  
     336, 337

- philography, 1, 5, 26, 31–33, 37, 193, 194, 213  
 Phoenician, 78, 112, 196, 197, 385  
 phonophore, 675  
 photocomposition, 634, 635  
 pine tree, 120, 131  
 pinyin, 68, 73, 136, 145–147, 330  
 Portuguese, 69  
 proto-Hebraic, 385  
 Punjabi, 52, 322, 324, 325, 337, 339, 340, 343–345, 355, 358, 360–363, 427, 443  
  
 Quechua, 506, 511, 534, 540, 544, 551, 567, 578, 579  
  
 radical, 6, 22, 37, 38, 48, 94, 382, 641, 671–686, 708, 709, 714, 715, 722, 724, 801  
 Rapa Nui, 471, 472, 474, 475, 477, 478, 480, 481, 485, 486, 489, 493–496  
 rasm, 318, 321, 326, 327, 332  
 RDF, 268, 272–274, 276, 278, 286  
 Recogito, 289  
 representative glyph, 303  
 Robofont, 805, 806, 814  
 rongorongo, 471, 473–475, 477, 479–481, 483–485, 487–489, 491, 493, 495–497, 499, 543, 548  
 rose, 1  
 Rosetta Stone, 123, 193–197, 199–207, 209, 211, 213, 215, 217  
  
 Sans Forgetica, 102, 104, 105, 107  
 Sanskrit, 53, 419, 420, 423, 427, 428, 433, 443, 702–707, 711, 723  
 Saussure, Ferdinand de, 12, 96, 118, 158, 221, 694, 701  
 Schriftlinguistik, 2, 8, 16, 19–23, 25, 26, 30, 34, 36  
 scripto continua, 378  
 scriptology, 23, 28, 30, 35  
 scriptura continua, 243, 245, 250  
  
 semanticity, 671–673, 675, 677, 679, 681, 683–687  
 semilexicality, 177, 179, 181  
 Semitic, 59, 316, 377–379, 381–387, 389  
 sentence-final, 157, 160, 162, 170, 172–176, 182, 184, 185, 189  
 Shahmukhi, 324–326, 328, 331, 333, 337, 339, 340, 343, 345, 358, 360–363  
 Sharada, 327  
 SignWriting, 785, 788, 789  
 Sindhi, 322, 325–327, 331, 336, 337, 339, 340, 342–347, 349–353, 355, 358, 360–362, 364, 432, 443  
 sinograph, 135–153, 155, 671–675, 677–681, 686  
 slang, 166, 401  
 Sogdian, 330  
 Sorani, 322, 328–330, 337, 339, 343–345, 347, 355, 358, 359, 361–363  
 SPARQL, 274, 289, 293  
 SVG, 267, 272, 273  
 syncretism, 85, 86  
  
 TEI, 288, 312  
 textel, 302  
 textogram, 82, 83, 87  
 texton, 305  
 TikTok, 166, 450  
 Tiwanaku, 501–506, 509–512, 516–518, 520–526, 532, 552, 554, 556–558, 561, 564, 565  
 Tlön, 121–123  
 Tolkien, J.R.R., 639, 642–644, 648, 651  
 toposyntax, 82, 87  
 Tschernochvostoff, Georg, 240  
 Turkic, 254, 316, 322, 330–332, 337

- Twitter, 162, 165–168, 170–176,  
183, 185, 411, 450, 451
- Twombly, Cy, 115, 117–119, 126
- Typannot, 781–783, 785, 787, 789,  
791–793, 795, 797, 799,  
801, 803, 805–815
- typème, 301
- typification, 151, 153
- typochar, 301
- typoglyph, 301
- typograph, 13, 18, 25, 26, 32, 79,  
95, 101, 102, 104, 113, 114,  
147, 149, 151, 244, 251,  
261, 262, 301, 323, 402,  
422, 611, 615, 616, 625–  
631, 634, 636, 653, 658,  
667, 668, 706, 781–783,  
785, 787, 789, 791–795,  
797–799, 801, 803–805,  
807–809, 811, 813–815
- UEFA Nations League, 253, 254
- UNESCO, 443, 444, 448
- Unicode, 79–81, 163, 164, 166, 184,  
232, 266, 270, 278, 279,  
289, 299–307, 309–311,  
313, 315, 317–322, 335,  
336, 354, 355, 439, 442,  
444, 449, 451, 452, 457–  
459, 461, 651, 675, 677,  
679, 695, 808, 813
- unqu, 76, 306, 503, 508, 527–534,  
542, 546, 549, 560–562,  
566, 567, 571–575, 581–  
583
- Uqbar, 121–123
- Urdu, 317, 318, 322, 323, 325–328,  
331, 333, 336, 337, 339,  
340, 343–346, 348–353,  
355, 358, 361, 362, 364,  
443
- Utopia, 639, 641, 642
- Uyghur, 317, 320–322, 328, 330–  
332, 337, 343–346, 348–  
355, 358, 361, 362, 364,  
702
- Uzbek, 320, 322, 330
- virama, 427, 428, 431
- Voynich manuscript, 123, 134, 213
- Wari, 501–507, 509–512, 516–518,  
520–523, 525, 526, 532,  
533, 537, 554, 556–565,  
574–577, 582
- Wikipedia, 62, 331, 337, 339, 340,  
349, 400, 679
- word order, 161, 172, 175, 176, 210,  
211, 741
- WordNet, 279
- writing system, 1, 2, 4–7, 10, 12,  
13, 16–19, 25, 27–30, 36,  
46, 49, 51, 62–64, 67,  
68, 70, 77, 79, 96, 112,  
123, 136, 150, 194, 213,  
219, 220, 237, 239, 243,  
253–256, 258, 260, 262,  
265, 312, 316–318, 320–  
330, 333, 335, 336, 354,  
377–380, 384, 385, 387,  
402, 419–423, 430, 437–  
445, 448, 449, 451, 452,  
458, 461, 462, 502, 512,  
513, 527, 546, 551, 571,  
639–651, 671, 673, 674,  
696, 698–705, 722, 723,  
741, 772
- XML, 288, 312, 339
- Yiddish, 317, 612, 618
- Zapotec, 450, 451
- Zipf's law, 471, 474, 475, 483, 497,  
740