

GRAPHOLINGUISTICS AND ITS APPLICATIONS

Grapholinguistics in the 21st Century—2024

/gʁafematik/
Proceedings

October 23–25, 2024, Venice, Italy
Yannis Haralambous (Ed.)

Part I

Fluxus Editions

Grapholinguistics and Its Applications 11

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/

October 23–25, 2024 – Venice, Italy

Proceedings

Part I

Fluxus Editions

Yannis Haralambous (Ed.). 2026. *Grapholinguistics in the 21st Century. October 23–25, 2024. Proceedings* (Grapholinguistics and Its Applications, Vol. 11). Brest: Fluxus Editions.

This title can be downloaded at:

<http://fluxus-editions.fr/gla11.php>

© 2026, The respective authors

Published under the Creative Commons Attribution 4.0 License

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

ISBN: 978-2-487055-10-0

e-ISBN: 978-2-487055-11-7

ISSN: 2681-8566

e-ISSN: 2534-5192

Cover illustration: Votive offering from the Sanctuary of Asklepios on the south slope of the Athenian Acropolis (350–300 BC, Acropolis Museum, EAM 15244): a marble face, its eyes inlaid in coloured stone, set into a niche carved in a limestone pillar. An inscription, still legible when the pillar was excavated in 1876, read: “ὑπὲρ τῆς γυναικὸς // εὐξάμενος // Πραξίας Ἀσκληπιῶ” (on behalf of his wife, having made a vow, Praxias [dedicated this] to Asklepios) (*IG* II² 4372). She was suffering, we may assume, from a disease of the eyes. Praxias could have contented himself with writing “I thank the god for healing my wife’s eyes.” Instead, he inscribed the rock with her face. Here writing is pushed to an extreme: the stone is the page, and the face—placed, framed, and addressed to a reader, the god—is the grapheme. Between the sentence carved in letters and the face carved in marble there is no break, only a continuum of graphic acts. It is this continuum, in all its breadth, that grapholinguistics takes as its object.

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: LuaL^AT_EX, bibl_{at}ex+biber (authoryear-icomp style), xindex

Fluxus Editions

38 rue Émile Zola

29200 Brest, France

www.fluxus-editions.fr

Dépôt légal: septembre 2026

κγζκς

Table of Contents

<i>Preface</i>	ix
<i>List of Participants at the Grapholinguistics in the 21st Century 2024 Conference</i>	xi

PART I

FLORIAN COULMAS. – The Linguistic Landscape: What we can learn from it	1
PHILIPPA M. STEELE. – The Visual-Cultural Fit of Writing Systems	15
MARY C. DYSON & DAVID BŘEZINA. – Do we process characters as shapes, nothing special?	47
DENNIS REISLOH & TATJANA SCHEFFLER. – On the linguistic status of emojis	65
DAVID A. ROBERTS. – Multilinear featurality in Roman script African orthographies	95
JONAS ROMSTADT. – Looking for the crucial point(s) – Forms and functions of punctuation units in handwritten German school exams	123
ALEXANDRE SOLCÀ. – Un court manuel du minoen proto-gréco- anatolien	159
ADELI V. BLOCK. – Inscribing Multilingualism: Lifemaps and “Trans- scripting” Among Moroccan Amazigh Youth	163
SPYROS ARMOSTIS & MARILENA KARYOLEMOU. – Creating a writ- ing system for an endangered language: the graphisation of Cypriot Arabic	181

JOSEPH DICHY. – Written languages and the Mental Lexicon: a crucial aspect of grapholinguistics, exemplified by agglutinative word-form recognition in Arabic and Biblical Hebrew	209
GUY BOURSIER. – Writing Egyptian Arabic: the case of a ‘dialectal’ Wikipedia	229
SINA FAKOUR. – Reviving Linear Elamite (2300–1880 BCE), The oldest known phonetic writing system.	247
YANNIS HARALAMBOUS. – Masoretic Graphemics and Graphetics: Challenges and Solutions	297
BEHNOOSH NAMDARZADEH & NICOLAS BALLIER. – Audio LLM subtokens as encapsulated knowledge: the case of Persian subtoken graphemic representations in Whisper	327
SAIYA L. KARAMALI. – The Developing Orthographic Conventions of Roman Hindi-Urdu	361

PART II

DONALD E. KNUTH, interviewed by YANNIS HARALAMBOUS. – The Grapholinguistic Model of T _E X, followed by an “All Questions Answered” Session	393
JAN KUČERA. – Tamil Orthography Reforms	417
VASANTA DUGGIRALA, RAJU SURAMPUDI BAPI & SOHAM G. KORADE. – Graphic Complexity and Orthographic Neighbors: Observations based on Telugu Writing System	443
PAUL D. UEDA. – Characters, Romanization, and Anti-Colonialism: Historical Development of Taigi and Vietnamese	469
HELEN MAGOWAN. – Towards reading in four dimensions	489
JAMES MYERS. – Synchronic and diachronic ordering in Chinese character form patterns	507
JANET GANON-DAVEY. – Beyond the alphabet, beyond the binary, maybe beyond “Chinese”? Exploring the emergence of gender-inclusive Chinese pronouns	533
NOAH HERMALIN. – The Efficiency of Spelling-Sound and Spelling-Meaning Mappings in Two Logographic Writing Systems . . .	551

GORDON G. BERTHIN & ROBERT SCHOCH. – The Antiquity and Timeframe of the Rongorongo Script on Rapa Nui (Easter Island)	597
ANTONIO PERRI. – Rethinking Unicode: how to digitally encode non-linear written artefacts? A tentative encoding of Codex Mendoza, folio 2r	633
ARVIND IYENGAR. – Braille as writing: Evidence and lessons for grapholinguistic theory	661
MICHAEL FILHOL. – AZVD as a Sign Language writing system proxy, and the potential evolution	689
MANGAS HECTOR AFONSO & PETER BILAK. – Enhancing Distance Reading for Low Vision: A Reading Acuity Experiment on Letter Width	709
TITUS NEMETH & MARTIN TIEFENTHALER. – Prima — An innovative type, designed for learning to read and write	713
SANTHOSH THOTTINGAL. – Parametric type design in the era of variable and color fonts	757
BOTH PARTS	
<i>Index</i>	α'

List of Participants

at the

Grapholinguistics in the 21st Century 2024

Conference

Ahlberg, Aija K.
Armostis, Spyros
Ashourinia, Kaveh
Auda, Iris
Avelino Corral, Esteban
Bacon, Bennett
Behnoosh, Namdarzadeh
Behrendt, Doris
Berthin, Gordon
Bianchini, Claudia Savina
Birk, Elisabeth
Block, Adeli V.
Bodor, Cristian
Boursier, Guy-Marie
Březina, David
Halverson, Colin
Cacciafoco, Perono
Candida, Fabiana
Capo, Daniele
Carlotti, Agnese
Cheminant, Anne-Claire
Circluede, Camille
Cockrum, Paul
Coulmas, Florian
D'Angelo, Antonietta
Danet, Claire
Dichy, Joseph
Dunlavey, Nicholas
Dürst, Martin Jakob
Dyson, Mary C.
Elta di Rodeano, Sveva
Fakour, Sina
Filhol, Michael
Fiona, Ross
Galli, Giulio
Ganon-Davey, Janet
Gardner, William L.

Gaultney, James V.
Gnanadesikan, Amalia
Groshi, Artur
Grzelec, Malgorzata
Guidotti, Marta
Guthrie, Tilly
Haralambous, Yannis
Harbour, Daniel
Hermalin, Noah M.
Hockett, Sara
Home, Eleanor
Honda, Keisuke
Huot-Marchand, Thomas
Iyengar, Arvind
Jindai, Keisuke
Joyce, Terry
Karamali, Saiya Layana
Khoury, Gretta
Khoury, Yara
Knuth, Donald E.
Krelaus, Leonie
Kučera, Jan
Lerbs, Meyer-Lothar
Magowan, Helen
Mastoridis, Klimis
Maurizio, Isabella
Meletis, Dimitrios
Mihaylov, Dimitri
Monaco, Sergio
Myers, James Tomlinson
Nemeth, Titus
Nevins, Andrew
Osgardy, David J.
Payne, Annick
Perondi, Luciano
Perri, Antonio
Polo, Ludovica

Raju S., Bapi
Roberts, David A.
Romstadt, Jonas
Schoch, Robert M.
Share, David
Shomer, Michal
Steele, Philippa M.
Strobel, Fabian

Szabolcs, Vátány
Terada, Yusuke
Thottingal, Santhosh R.
Tiefenthaler, Martin
Ugray, Gábor Laszlo
Vasanta, Duggirala
Verna, Didier
Xu, Duoduo

The Linguistic Landscape: What We Can Learn From It

Florian Coulmas

Abstract. In the course of the last two decades, the linguistic landscape, that is, the display of written language in public places, has become an object of investigation, which is now recognised as a branch of Sociolinguistics. This is remarkable because since its inception, Sociolinguistics, much like structural linguistics, has always had an oral bias and paid relatively little attention to writing, which was often seen as a mere representation of speech. In fact, written language is a branch of human communication with great social significance. In this lecture, I will show that linguistic landscaping is a political issue with many economic and cultural aspects, which upon inspection reveal much about how a society is constituted. At the same time, the linguistic landscape is a repository of history that is often contested. Using examples from various contexts in Asia and Europe, I will show how and why language displayed in public texts may become and often is subject to legal restrictions designed to proscribe or, in the contrary, promote the visibility of languages in the public domain. Language, in writing as well as in speech, is always embedded in social practice. This is only pointing out the obvious. However, I will try to show that the linguistic landscape also makes for a good starting point with regard to foreign language education.

1. Introduction

Once we have opened our eyes to it, the linguistic landscape (LL) presents itself in such a variety that every student of it must grapple with the question of how best to conceptualize and order it and how to assign it a position in our understanding of language and society. In the initial phase when Sociolinguistics gained recognition as a proper field of study, research was mainly oriented to spoken language and the many varieties which earlier investigations of (normative) grammar had largely ignored. About the relationship between language and society, written language has, however, as much to say as speech. Since early studies near the end of the last century, for instance by Landry and Bourhis, who defined the linguistic landscape as “the visibility and

Florian Coulmas  0000-0002-6028-6464
IN-EAST Institute of East Asian Studies & University of Duisburg-Essen

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 1–14. <https://doi.org/10.36824/2024-graf-coul>
ISBN: 978-2-487055-10-0, e-ISSN: 978-2-487055-11-7

salience of languages on public and commercial signs in a given territory or region" (1997: 23), the field has expanded steadily. My view is that the field is still evolving with new definitions being proposed, as new research projects are undertaken.

Rather than introduce yet another definition, I want to start here by walking you through a series of examples that may give you a *tour d'horizon* of the field. It is by no means representative, let alone complete, but may illustrate the multifariousness of life with the written word in the public sphere.

Linguistic landscape is a way of looking at things or, if you like, a lens through which to see writing, and this lens reveals many different facets. For instance, aesthetic: is it beautiful or ugly? Linguistic: are words and grammar used correctly? Are there many mistakes? Geolinguistic: is the LL what you would expect of the region/country where you are? Sociolinguistic: What languages and varieties are being used? Political: Do messages on display carry political contents? Is the script ideologically loaded? Economic: Who pays for, and who benefits of, inscriptions on public signs? Historical: Since when have signs been where they are now and why? And so on. Many other questions could be asked. But let's get a bit more specific.

2. A Guided Tour Through the LL

My proposal is to consider LL research as a branch of Sociolinguistics while drawing on other theories wherever that is useful.

Linguistic landscapes are part of a visibility regime. The eight *wh*-questions—*who*, *where*, *when*, *what*, *which* language(s), to *whom*, *why*, with *whose* permission—are a first step toward a systematic analysis of such a regime.

LLs consist of linguistic expressions of various kinds exhibited in public spaces that are initially partitioned spontaneously. With growing complexity of urban environments legal regulations evolve that maintain divisions and monitor them more or less strictly. Where norms and regulations are established *visibility regimes* come into existence. Important as they are, *LLs are only one part of visibility regimes*. The display of religious symbols, flags, dress codes (veil in public), seating arrangements (at school, in parliament), art styles, architecture, license plates, among others are also included. Considering all of them, it is important to pay attention to the *boundaries between the visible and the invisible* by means of which social figurations manifest themselves as a visual order. To put it differently, LLs are part of a visibility regime that reflects government regimes. What is visible and what invisible may be an aesthetic question, but in many cases it is politically loaded.

The boundaries between the visible and the invisible are social facts. In relation to LL they refer, for instance, to language status, that is, national, official, minority, or foreign. Are there any restrictions on displaying a language in public? Do such restrictions become apparent upon comparing a particular LL now with the same LL two, three, four decades ago?

We can look at the boundaries between the visible and the invisible of a LL from the point of view of *appearance discrimination*, e.g., the teaching of Chinese languages in schools and Chinese on public signage was banned in Indonesia from 1966 to 1998.

Under dictator Francisco Franco (1939-1975) the Catalan language was banned from public spaces, even from gravestones, in Spain where Spanish was made the sole language of public life.

Bilingual street signs vandalized in many places, Ulster, Carinthia, Austria.

Foreign language street signs as *tokens of recognition*, e.g., Japanese, Arabic in Düsseldorf; metro station signs in 4 languages in Tokyo.

Hierarchies in bi-/multilingual street signs, e.g., in Quebec: “The Office Québécois de la langue française considers any message displayed in a place accessible to the general public as “public signage”.”

The general rule is that public signage and commercial advertising in Québec must be in French. Another language may also be used alongside French, provided that French is ‘markedly predominant’.¹

Making (in)visible is a way of exercising power.

As a theoretical embedding I will use 8 Wh-questions, as guideline as follows:

1. *Who* writes
2. *Where*
3. *When*
4. *What*
5. Using *which* language(s)
6. addressing *Whom*
7. *Why*, to what end
8. with *Whose* permission.

3. Examples

With these questions in mind, let us go through a round of linguistic landscape clippings.

1. L’affichage public et la publicité commerciale doivent se faire en français. Ils peuvent également être faits à la fois en français et dans une autre langue pourvu que le français y figure de façon nettement prédominante. Charte de la langue française. <https://www.legisquebec.gouv.qc.ca/fr/document/lc/C-11>

3.1. Kanji of the year 2022



Who: Kiyonori Mori, the head priest of Kiyomizu Temple, who also served as a director of the Japan Kanji Proficiency Test Association. *Where:* Kiyomizu-dera, Kyoto. *When:* 12/2022. *What:* the Chinese character of the year: 戦 *sen, ikusa* which means war, battle. In *which* language: Japanese/Chinese. To *whom:* the general public. *Why:* ???

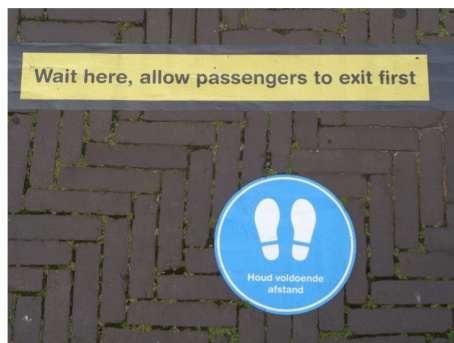
With *Whose* permission: Kiyomizu Tempel and Japan Kanji Proficiency Test Association.

3.2. Covid-19 The Hague



Who: private person. *Where:* Amsterdam. *When:* 12/2022. *What:* Covid-19. *Which* language: indefinite. To *Whom:* everyone. *Why:* ??? *Whose* permission: without.

3.3. Covid, passengers



Who: city administration. *Where:* The Hague. *When:* 2022. *What:* instruction to guide behaviour. *Which language:* English, Dutch. To *Whom:* passengers *Why:* health protection. *Whose permission:* ???

3.4. Covid African Artists, De Fred



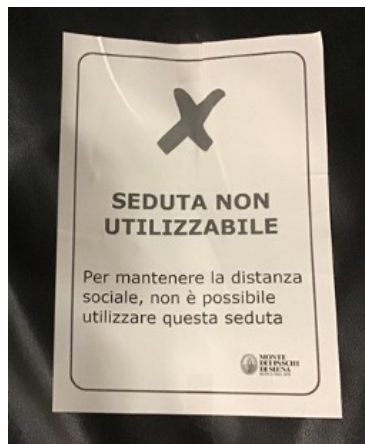
Who: African artists. *Where:* Berlin, Cyber space. *When:* 2022. *What:* publicity. *What language:* English. To *whom:* public. To *what end:* ??? *Whose permission:* ???

3.5. The Hague



Who: Neighbourhood. *Where:* The Hague. *When:* 2022. *What:* safety instruction. *Which language:* Dutch. *To whom:* public. *Why:* fight pandemic. *Whose permission:* Neighbourhood association.

3.6. Distanza sociale—new expressions



To maintain social distance
this seat cannot be used.



Wait here. Observe
the safety distance.

Who: A bank, a shop. *Where:* Tuscany. *When:* 06/2022. *What:* health safety. *Which language:* Italian. *To whom:* customers. *Why:* to fight Covid. *Whose permission:* Bank, shop owner.

3.7. Quarantine, St. Luke's



Who: St. Luke's Hospital. *Where:* Tokyo. *What:* Encouragement. *Which language:* English, Japanese. *To whom:* staff, patients. *Why:* to boost morale. *Whose permission:* Hospital.

3.8. How to confront the pandemic in Tokyo

English first, Japanese second?



Who: Ministry of Health. *Where:* Tokyo airport. *When:* 2023. *What:* instruction. *Which language:* Japanese, English, Chinese, Korean. *To whom:* passengers. *To what end:* Instruction. *Whose permission:* Japanese government.

3.9. Non siamo pacchi

Demonstration in Venice features trade union's demands, which throws some light on freedom of expression in Italy.



Who: kindergarten teachers. *Where:* Venice. *When:* 2020. *What:* protest. *Which language:* Italian. *To whom:* government. *Why:* to get a raise. *Whose permission:* ???

3.10. Twelve socialist core values

How the public sphere is governed, in China compared to other countries. From left to right: Prosperity/Democracy, Civility/Harmony, Freedom/Equality, Justice/Rule of law, Patriotism/Dedication, (not visible:) Integrity/Friendship.



Who: Tongji University. *Where:* Shanghai. *When:* 11/2019. *What:* Socialist values. *What language:* Chinese. *To whom:* Students *Why:* political education. *Whose permission:* State.

3.11. Street sign in Jerusalem



Who: Municipal administration. *Where:* Jerusalem. *When:* 10/2022. *What:* Street sign. *Which languages:* Hebrew, Arabic, English. *To whom:* public. *Why:* to represent all population groups. *Whose permission:* the city's.

3.12. City of tolerance, Lisbon



Who: City of Lisbon. *Where:* Portugal. *When:* 2016. *What:* public commitment. *Which language(s):* many. *To whom:* citizens and foreigners. *Why:* to promote tolerance symbolized by multilingualism.² *Whose permission:* the city's.

2. *Editor's note:* Notice that while Portuguese and Italian use classical 19th-century-like type, Greek and Hebrew are set in Helvetica, and Arabic in Geezah, a modern computer-like typeface. All languages are equal, but some share the esthetics of the environment while others are represented in a modern and cold way.

3.13. “Without a valid ticket”

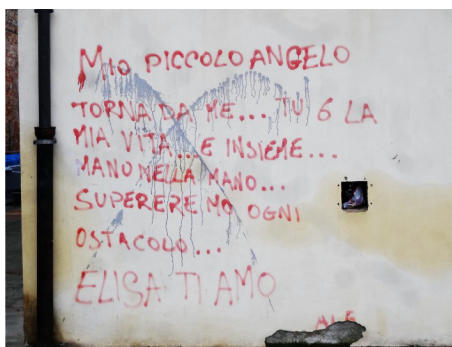
Where multilingualism in the public sphere is important.



Who: Public transport. *Where:* Berlin. *When:* 2023. *What:* warning to free riders. *Which language(s):* many. To *whom:* citizens and foreigners. *Why:* to make them pay the fare. *Whose permission:* the city's.

3.14. Mio piccolo angelo

Graffiti with an addressee in Venice: My little angel come back to me ... you are my life ... and together ... hand in hand ... we will overcome any obstacle ... Elisa I love you (Photo 09/2016).



Who writes? ALE. *Where:* Venice. *When:* ?? *What:* a plea, declaration of love. In *what* language: Italian.³ To *whom:* Elisa. *Why:* to get her back. With *whose* permission: without.

And who crossed it out?

3. *Editor's note:* Note the rebus-like style of writing “sei” (“you are”) as a digit “6,” also written “sei”.

3.15. Ayasofya からすみ専門店

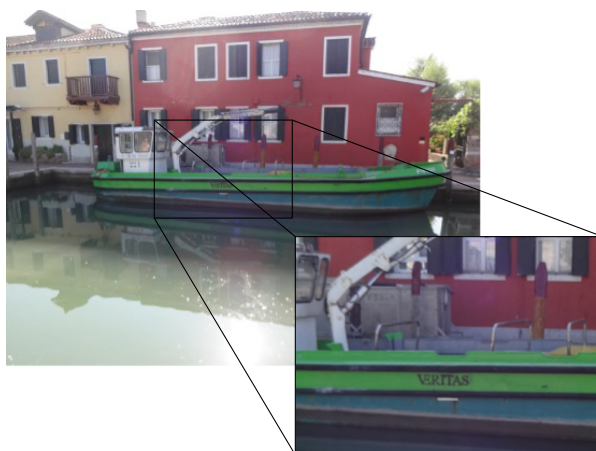
Solving a riddle: Why Japanese ads for smoked roe and beds covers in shop in Istanbul?



Who: shop owner. *Where:* Istanbul. *When:* 09/2014. *What:* ad, decoration. *Which language:* Turkish, English, Japanese. *To whom:* tourists, public. *Why:* commerce. *With whose permission:* ???

3.16. The truth: VERITAS

(VERITAS = Veneziana Energia Risorse Idriche Territorio Ambiente Servizi)



3.17. Signs of the times

In Venice, the truth is still expressed in the local language. In Tokyo, a city not known for the multilingual skills of its native population but which currently experiences a high rate of immigration, the truth can now be expressed in many languages, at least when it comes to crossing the street.



3.18. 大文字 festival

About Japanese culture and the importance of writing in it: Chinese character 大 *dai* 'big' cut into a mountain outside Kyoto and lit at the Daimonji Festival in August. The purpose of the bonfires of the festival is to guide the spirits of the ancestors back to heaven.



4. Conclusion

I hope I have given you an idea or at least a presentiment of how full of surprises and unexpected features waiting to be discovered and analysed the linguistic landscape holds.

LL research can be approached from different disciplinary angles. To begin with, since nowadays every LL is located in a state, you may want to ask whether there are any legal norms regulating the public display of writing. To take an obvious example, graffiti is widely considered a crime that exposes the writer/artist to civil liability for the damage to the defaced property. This is why the last one of my seven *wb*-questions refers to permission. Are there any such laws? If so, what is the contents? Does it contain a definition of “graffiti”? How severe is the penalty? Have there been any interesting precedents? These are just some of the questions to be pursued in this connection.

Example: The Indonesian Presidential Instruction No. 14/1967 (Inpress No. 14/1967) on Chinese Religion, Beliefs, and Traditions effectively banned any Chinese literature and cultures in Indonesia, including the prohibition of displaying Chinese characters in public.

Politically motivated renaming of street names, e.g., a segment of the street where the Soviet embassy was located in Washington DC, in 1984 was renamed Andrei Sakharov Plaza, after the jailed physicist and Soviet dissident.

Quite recently, in December 2022, the Nachtigalplatz in Berlin, so named after the German empire’s commissioner for West Africa who played an important role in the German colonisation of Togo, Cameroon and Namibia in the 1870s, has been renamed Manga Bell Platz in memory of Rudolf and Emily Duala Manga Bell, a king and queen of Duala in Cameroon who fought against German colonialism. The examples of taking issue with historical legacies in street names are many, and often they come along with fierce debates between different interest groups. In the event, historical memory wasn’t eradicated, but the question what should be remembered was answered differently.

Next, *education*. The socialist core values publicly exhibited in Chinese cities are an example of an authority claiming the linguistic landscape for purposes of education. The political motivation is very obvious here, but when we start asking why streets are named as they are, we will find a variety of reasons, political ones among them. Since the appreciation of politicians and political events is historically contingent, evading the necessity or desire of renaming streets is also an option. In Tokyo, for example, most streets have no name. One’s address is just a number code. Or consider New York city. 5th Avenue, surely one of the best-known streets of the US if not the world, will never have to be renamed.

There are beautiful street names, such as *Love Lane* in London, frightful ones, such as *Calle dei assassini*, Assassin Street, in Venice, and run-of-

the-mill street names, such as *Main Street* or *Rue de l'Église*, Church Street, the most common street name in France. Just names, you may want to say, but it actually matters. Street names are residents' addresses, and there are popular and unpopular addresses. Remember Monopoly, the ultimate capitalist parlour game? Streets are ordered there for value, Park Avenue being the most expensive. Research has shown that street names affect real estate prices. There are hence also *economic* aspects of the linguistic landscape—among others.

The lesson that I draw from these observations is this. The Linguistic Landscape is all around us. Wherever we go, it is a multifarious configuration of languages, authors, readerships, passers-by, observers, supervisors, censors driven by political, social, educational, cultural, and economic motives guided by and transgressing standards of writing in public, all of which can be studied with equal benefit for our understanding of the ever-changing dynamics of language in society. Because of this *multiformity*, it is not surprising that there is *no generally agreed on definition of LL*, nor is there one theory that can direct its investigation. The linguistic landscape is an artefact that keeps changing as society evolves. There are always new things to discover in the linguistic landscape, such as this Chinese character addressed to spirits to find the way to eternity.

The Visual-Cultural Fit of Writing Systems

Philippa M. Steele

Abstract. Writing systems have sometimes been evaluated in terms of their “linguistic fit” to the language written in them, a concept that has pervaded a good deal of literature on writing typology. Here I argue, building on the idea that writing systems also have a broader sociolinguistic fit (Meletis 2020), that the visual-cultural fit of a writing system in its sociolinguistic context should also be considered. This reveals an important principle of writing system development, namely visual distinctiveness. Following an introduction to the concepts discussed in the paper (§1), the ancient writing systems known as Linear B and the Cypriot Syllabary are used as a comparative case study illustrating that linguistic and visual-cultural fit need not and often do not have a strong correlation (§2). Different types of visual distinctiveness in writing are then presented, with a particular focus on systems used for minoritised languages and groups in the modern day, where visual distinctiveness may be particularly important to the community of users (§3). Finally the conclusion acknowledges further practical problems that affect minority writing systems, particularly digital disadvantage and issues related to dissemination (§4). The paper demonstrates that visual aspects of writing deserve further attention in and beyond grapholinguistic study.

1. Introducing “Fit” and “Vitality”

The idea that a writing system can be well-suited (or ill-suited) to a language, or that a language can have an ‘optimal orthography’ (Rogers 1995), has persisted as a tool for thinking about graphic representations of language. The concept of *linguistic fit* can be thought of as a way of assessing the “quality” of a writing system’s semiotic structures (Meletis 2020, 231), or the effectiveness with which linguistic information can be encoded by writers and decoded by readers, with potential for wider ramifications among its users. Real-world applications could be considered to be a separate area of enquiry (cf. the separate categories of “processing fit” and “sociolinguistic fit” suggested in Meletis 2020). To take

Philippa M. Steele  0000-0003-3109-265X
University of Cambridge, Magdalene College, Cambridge CB3 0AG, UK
E-mail: pms45@cam.ac.uk

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 15–46. <https://doi.org/10.36824/2024-graf-stee>
ISBN: 978-2-487055-10-0, e-ISBN: 978-2-487055-11-7

one brief example, the use of an antiquated orthographic system for *Tibetan*, with a considerable mis-match between written and modern spoken forms, has been demonstrably linked with learning difficulty and hence with low literacy rates (Tournadre and Gsang-bdag-rdo-rje 2003, and see recently Schmidt 2024): this is largely a result of political insistence on linguistic and so also written unity in a linguistically diverse state (see Roche 2020, 78; also Roche 2024, 70-75), although the single orthographic system at the same time can be used as a sort of interscript between different language varieties, none of which are well represented by it. Even such a brief example shows that writing systems exist within and interact with messy networks of linguistic and social structures, and their operational effectiveness can only be understood in that broader context.

Thus, although linguistic fit is a *descriptive* tool of grapholinguistics (e.g., Meletis 2020, 231), a contextualised assessment of a given writing system can have positive or negative associations that result in judgements of value or effectiveness (for various takes on this, see Katz and Frost 1992, Seidenberg 2011, Frost 2012, Perfetti and Harris 2013). It would not be unreasonable to suggest that the application of this grapholinguistic theory is a way of testing of its own validity. The present paper takes the position that the application of grapholinguistic theory is of great importance, particularly in this case because it may have some bearing on the global language crisis that today sees more than half the world's living languages in danger of imminent loss within the next century (6,769 languages of which 60% are on the scale between "threatened" and "nearly extinct" according to Glottolog at time of writing: <https://glottolog.org/>). Since writing can play an important role in the fight against *language endangerment* and loss (see Shah and Brenzinger 2021), understanding how writing supports language should be a priority.

Perhaps the most important point to make at the outset is that it does not necessarily follow that a writing system with a good linguistic fit will be associated with high literacy rates, widespread use, or notable longevity (i.e., possible indicators or outcomes of vitality in the sense presented below). In other words, there is not necessarily a straightforward relationship between readability/usability and acceptability, or between linguistic fit and implementation (cf. Dawson 1989, Cahill 2018). The development of orthographies (very often adaptations of the Latin alphabet) for previously unwritten, minoritised languages is a case in point (see Grenoble and Whaley 2005, 102-159; Lüpke 2011; Shah and Brenzinger 2021), as case studies show that devising the orthography is only the beginning of the process, while community consultation and acceptance are ultimately the deciding factors in the system's take-up (see the papers in Jones and Mooney 2017 for a range of examples). Many such orthographies have suffered from limited take-up and/or have gone

out of use quickly, meaning they have not been effective supports for the language they were intended to write. The present paper is part of an ongoing programme of research to address this issue by considering the factors that contribute to the vitality of a writing system, thereby making it a more effective support for the language it serves. To an extent, this means it is important to consider writing beyond the narrower study of mapping linguistic units to graphic units, since contextual factors will play a significant role in such outcomes, and, overall, an integrated approach is necessary.

The concept of linguistic fit can itself refer to several aspects of language, not only phonological representation (i.e., whether all of the language's phonemes are represented). The issue of structural suitability, for example, is also crucial to the choice of writing system type, and as Baroni emphasised some time ago, an alphabetic-phonological principle is not necessarily the best fit for many languages (Baroni 2011). Features related to the phonological repertoire, syllabification patterns, and morphological structure of a language, as well as others such as tonality and word length, require a system whose structure can accommodate them efficiently (see also, briefly, Meletis 2020, 234), whether by total graphic representation or through orthographic strategies. For instance, the *Japanese* syllabic kana systems (hiragana and katakana, with V and CV signs) could be said to be well suited to the language's syllable or mora structure and especially the frequency of open syllables (Joyce and Meletis 2021, 267), although it is important also to consider the reality of overall orthographic complexity when combined use of all the different systems for Japanese is considered (Joyce and Masuda 2019). These are language-specific concerns related to structure; I will deliberately leave aside the question of whether particular kinds of structural encoding may or may not have universal cognitive benefits, where important research is ongoing (see for example a recent study on the syllable as a unit of learning in Arabic: Abu Ahmad and Share 2024) but not yet sufficiently established across writing system and language types to be applied universally.

Some of the properties that determine writing system efficiency are not intrinsic to the language and its typology but rather to contextual factors, as already noted (see also Cahill 2014). Because contextual factors are not, or not overtly, linguistic, they pose a greater challenge for a grapholinguistic framework that draws its structure from general linguistics or semiotics. Meletis has proposed a category of "sociocultural fit," starting out with the observation that writing systems can have indexicality since 'they can signal, for example, parts of their users' identities by being connoted with a specific group of people' (Meletis 2020, 339), and admitting that there is such a wide range of possible features that to some extent they have to be 'bundled' together (ibid., 340); he then groups his examples and discussion into the three categories of

Graphetics, Graphematics and Orthography, of which the first is the most overtly concerned with visual matters. The broadness of the socio-cultural categorisation is appropriate, since writing is affected by many factors arising from social and cultural context, but it could be argued that it is too unwieldy to apply usefully to individual cases with anything other than a broad brush. Particularly in order to strategise for the future, it is necessary to break down areas of sociocultural relevance and tackle them individually. That is in part the rationale for considering whether “*visual-cultural fit*” could be a helpful way of thinking about the visual aspects of writing and their relationship with issues of linguistic and cultural minoritisation and endangerment (see also recently Simpson 2024)—so the intention here is to make a small contribution to a much larger topic. A further motivating factor is the marginalisation within grapholinguistic research of some writing systems whose features are relatively difficult to describe within a linguistic framework, and may be better served by a framework that includes visual features. The latter is a long-term aim to which I hope this paper may make a contribution.

Theoretically, writing can be a powerful support to a marginalised and/or threatened language because it can make long-distance and digitised communication possible, enable documentation and the creation of new language materials, lead to an online presence, influence other areas of culture such as art and design—on top of the fundamental function of providing speakers with a visible manifestation of their language, which for some may equate to a validation of its very existence and value (see e.g., Cahill 2004). But the successful take-up of a writing system cannot be predicted simply from its linguistic fit. One choice, for example, is between an orthography of an existing majority system on the one hand and a distinct minority system, perhaps unique to the language, on the other (whether a heritage system or a newly created one). This choice, which is in large part visual (though it may also relate to other choices, such as structural ones), plays a significant role in a writing system’s vitality—and so should be at the forefront when deciding how to write down a minoritised, indigenous, or endangered language. The choice should also be contextualised, rather than relying on assumed universals as mentioned above (such as the syllable as an effective learning unit or the phonemic principle as supposedly democratising). Unseth, speaking about emergent scripts in Africa, offers three criteria necessary for the successful adoption and continued use of a writing system: community endorsement, a corpus of literature, and a schooling programme (Unseth 2011, 31), factors that we will return to in the final section.

The idea that a writing system can have its own vitality can be situated within a much broader framework of cultural endangerment since writing systems are different kinds of entities from language and rely on social praxis and materiality (see Steele 2023, 79–88). I am going to

use some examples from the ancient world to introduce this concept, since historical systems can more easily be assessed from a teleological perspective—that is to say, we know the answer as to whether the system survived long-term or not, and, depending on the quality of the evidence, we may be able to make a more detailed assessment of its vitality in different areas and periods of use. Being able to study long-term trajectories in historical writing systems is an advantage not shared by most minority writing systems today, since many are poorly documented and/or recent creations or codifications. Linguistic fit is not necessarily highly correlated with writing system vitality or positive outcomes, as far as historical examples show. The extent to which visual-cultural fit correlates with vitality is a central testing ground for this paper.

In order to evaluate vitality from a teleological perspective, I am going to use a very basic set of three indicators for historical writing systems:

- Longevity: How long was the writing system in use?
- Literacy: How many people could use the system, and over what range of social spectra?
- Domains of writing: What was the range of uses of the system?

A wider range of factors is typically considered when assessing the vitality of living languages according to scales such as that published by UNESCO in 2003 (<https://unesdoc.unesco.org/ark:/48223/pf0000183699>), GIDS (Fishman 1991), or LEI (Lee and Van Way 2016, 2018). However, many of the indicators used in such scales (e.g., numbers of users, age of users) cannot be assessed with regard to historical writing systems, and arguably writing systems may require a different set of indicators reflecting their practical, material, visual, and embodied properties (see Steele 2023, 81–88). The UNESCO, GIDS, and LEI scales are also specifically geared towards assessing the health or vitality of a language from a synchronic perspective, or in other words, its present likelihood of surviving in the future. For historical systems, the teleological perspective is perhaps more helpful, since one of the major pieces of information we have is whether the system survived, and for how long—hence the emphasis on diachronicity here. Nevertheless, there is a relationship between these two, as diachronic, long-term usage should theoretically entail synchronic vitality for at least some periods of use.

2. Historical Examples: Linear B and the Cypriot Syllabary

The writing systems of the Bronze Age Aegean are a family of scripts applied to several different languages or language varieties. Both Linear B and the Cypriot Syllabary have been deciphered and were used to write varieties of Greek (Mycenaean and Cypriot Greek, respectively);

the Cypriot Syllabary was also used for at least one, probably indigenous language of Cyprus, traditionally labelled Eteocypriot. Cretan Hieroglyphic and Linear A were used in Crete (Linear A also a little more widely) to write at least one language usually referred to as Minoan, now poorly understood to the extent that these writing systems are usually labelled as undeciphered. The situation is more complex than that, as there are good reasons for assuming that the sign values identified for Linear B can largely be read backwards onto Linear A (and probably also Cretan Hieroglyphic) as at least an approximation, bearing in mind the likelihood of a different phonological inventory in Minoan language(s) (Steele and Meißner 2017). Cypro-Minoan, the final member of the family, also remains undeciphered and was used for at least one poorly understood language of Late Bronze Age Cyprus. We will concentrate here on Linear B and the Cypriot Syllabary, whose sign values and orthographic rules are well understood. Both could be argued to constitute examples of poor linguistic fit.

Linear B was used to write the earliest attested variety of Greek, between the 15th and 13th centuries BCE. Before its decipherment, many of the features of an earlier stage of Greek could be reconstructed, such as labiovelar consonants that later merged with other sounds; the decipherment confirmed these reconstructions, and quickly progressed to an advanced understanding of its linguistic features and the way they were encoded in the Linear B syllabic script (for the most recent full account, see Thompson 2024 and Thompson and Meißner 2024). Although Linear B could be classed as logosyllabic, in practice the syllabic and logographic inventories are kept separate, even where the same signs are used (i.e., polyvalency): unlike in contemporary, nearby logosyllabic systems like Anatolian Hieroglyphs and Mesopotamian Cuneiform, logograms could not be used syntactically (i.e., to stand for a word, or spell one in conjunction with syllabograms, within a sentence) (Steele 2023, 39-54). Instead, logographic signs were relegated to specific slots in the layout of administrative documents, and are also identifiable by their co-occurrence with numerals and measurement signs. The linguistic fit of Linear B to Mycenaean Greek will therefore be assessed primarily on the basis of its syllabographic repertoire, although the logograms could be argued to aid reading and thus improve processing fit to a limited degree.

When studying the way in which Greek was written in Linear B, a number of deficiencies are quickly apparent:

1. Linear B encodes only open syllables: V, CV, and occasionally CCV. Problem: Greek has very frequent consonant clusters in all word positions and often has final consonants.
2. The Linear B consonant series does not distinguish voicing (except in one case, the d-series) or aspiration. Problem: In Greek of this period, voicing and aspiration are phonemic contrasts (e.g., /k/, /k^h/, /g/).

3. Linear B does not distinguish between /r/ and /l/. Problem: In Greek, these are phonemically distinct.
4. Linear B does not distinguish short and long vowels. Problem: In Greek of this period, vowel length is phonemic.

Problems 2–4 are essentially failures of phonological representation, and have been suggested to originate from the phonological features of an earlier language for which the “ancestor” writing system Linear A (and/or Cretan Hieroglyphic) was designed. Linear B would then be the result of writing Greek in Linear A without doing very much to redesign the system. A few of these issues could be written off as ones that do not impede the use of the system very much. For example, although vowel length is phonemic in Greek, the occurrence of minimal pairs (i.e., pairs of words identical in all other respects than the feature in question) is perhaps not frequent enough to cause regular confusion. Even in nominal morphology, there are a few categories where vowel length is the only distinguishing feature. On the restrictions of the consonant series it has been argued that, since /b/ is notoriously underrepresented across Indo-European languages (i.e., it is difficult to reconstruct an original /b/ in the proto-language), it could actually have been *more* efficient not to represent it with a separate series of signs (Palaima and Sikkenga 1999, 602). However, this argument does propose any motivation for the total lack of distinction of aspirated consonants, nor that of the more frequent voiced sounds /g/ and especially /g^w/ (which is often the source of later Greek /b/). Arguably, grouping phonemes together could have a beneficial effect on the size of the syllabary, since for every consonantal phoneme differentiated, a series of five signs (e.g., *ga*, *ge*, *gi*, *go*, *gu*) would have to be added to the repertoire (Meißner 2007, 97). However, if this is an economical principle, it is at the expense of the ability to differentiate between phonemes, with a possible result of ambiguity in some spellings.

Problem 1, on the other hand, is deeply embedded in structure. Linear B signs (V, CV, CCV) must always write sequences that end in a vowel, and so would suit a language with predominantly open syllables—but the Greek language has many closed syllables in all positions, including at the end of words. So, in order to write Greek in Linear B, orthographic strategies are needed for consonant clusters and word-final consonants (see Consani 2003). Linear B employs a complex but highly standardised solution to this problem:

- Word-final consonants are always omitted (the only exceptions being where one member of a final consonant cluster is written, as in *wa-na-ka* for /wanaks/).
- Some consonant clusters are written with partial spelling (i.e., omit one of the consonants, as in *pe-ma* for /sperma/).

- Some consonant clusters are written with plene spelling (i.e., write both consonants but use a “dummy vowel” for the one that has no following vowel, as in *ko-no-so* for the place name Knossos).
- The second element of a diphthong is usually not written, although there are some exceptions (including the use of special signs *ai* and *rai*).

The rules for partial and plene spelling of consonant clusters are complex but fairly consistent overall, and depend on the position of the cluster and its composition in types of sound (see Thompson and Meißner 2024, 119–123). In word initial position, for example, full plene spelling is used if the first sound is a stop, but partial spelling is used in clusters with /s/ followed by a stop (where /s/ is omitted in writing). The choice of dummy vowel in cases of partial spelling is not entirely consistent but usually relies on syllabification (i.e., it is borrowed from the syllable to which the written consonant belongs).

The reason this complex set of solutions is so highly standardised is presumably that writing was a trained activity that was important for administrative professionals working in the Mycenaean centres, which is where almost all our evidence for Linear B writing survives in the form of bureaucratic clay documents. It could be argued that the necessity of complex orthographic rules would have increased the cognitive learning burden for the system, which can be an inhibiting factor for literacy, and so could be cited in support of restricted use of the system—essentially that it was used only in administrative settings. On the other hand, the set of orthographic choices was evidently designed with a practical principle in mind, since it ‘privileges the lexical access and the semantic identification of the lexemes, with simultaneous indifference for the rendition of the morpho-syntactic aspects of the utterances’ (Consani 2016, 96). Whether that is an advantage is more ambiguous. For example, the lack of differentiation of nominal endings, and so the underrepresentation of syntactic relationships, could perhaps be seen as a weakness for an administrative tradition that records a great deal of information about givers and receivers of various goods and materials. Judson (2022) has further demonstrated, based on variations in orthography at Pylos, that the writers were trained in a standard orthography for phonological sequences but did not have single “correct” spellings for words (even common ones) or even for morphological sequences, such that they would make decisions on how to spell a word based fundamentally on the phonological rules. That suggests in turn that the usability of the system hinged on standardisation at the level of phonological orthography, which may also help us to understand how it was first developed to write Greek, since the orthographic rules broadly hold true from the earliest to the latest attestations.

It remains to ask what effect the purported deficiencies of Linear B may have had on its vitality and its long-term use. While the dominant

Mycenaean administrative centre of Crete, Knossos, underwent destruction more than a century earlier, the last attested traces of Linear B come from mainland Greek sites such as Pylos and Thebes and date to the transition to the last phase of the Late Bronze Age (Late Helladic III B-C): around 1200 BCE. The clay documents written in Linear B only survive because they were baked in fires that destroyed the administrative complexes within which they had been agents of bureaucracy. It is impossible to say why or how those buildings were destroyed, but the old narratives of conflict (an invasion by outside groups) have given way to more nuanced attempts at explaining what was clearly a time of significant economic upheaval and social change. We can see from the archaeological record that some sites were abandoned, people were uprooted and probably moved to live in other areas, and the telltale signs of Mycenaean centralised administrations disappeared. The usual explanation is that there was no need for Linear B any more once the administrations no longer existed, and pre-existing social order had been overturned (perhaps resulting in the promotion of Mycenaean *qa-si-re-we*, some sort of minor official, to something more akin to the Homeric and later sense of βασιλεῖς: Morpurgo Davies 1979). This presupposes that Linear B was not used outside of the centralised administrations; however, we have insufficient evidence to decide this question. Might there have been people who used Linear B in other domains of writing (as believed by many Mycenologists) and carried on writing it in other contexts even after the fall of the administrations? In a climate where the archaeological survival of other potential writing materials like wood, papyrus, and parchment is unlikely, there is no secure basis for arguing one way or the other—this is an issue I have written about at length elsewhere and I will not rehearse the whole discussion here (Steele 2023, 110–125).

For Consani, the Linear B writing system was good *enough* for its administrative purposes, particularly bearing in mind the concomitant use of sematography for additional information—as long as we assume that the system was only used within the restricted contexts of literacy that are reflected in the surviving archaeological record (Consani 2016). If writing did occur in other domains, such as private or political letters or trading contexts, on the other hand, might writers have been hampered by the features of the script or even devised different orthographic rules? Perhaps it is worth asking the question, but there is a total dearth of evidence to answer it. However, a comparison with ancient Cyprus will be instructive: while its syllabic writing system is structurally very similar to Linear B, the evidence for literacy points in a very different direction.

Just like in Crete, in Cyprus, we have a situation where an undeciphered writing tradition is succeeded by one we can read, which was largely in Greek. Writing is first attested in Cyprus in the early part of the Late Bronze Age in a system labelled Cypro-Minoan by Arthur Evans because of its obvious link with the Cretan scripts. The small num-

ber of Bronze Age inscriptions are very diverse and mostly very short, and whatever language(s) they were written in remain unknown. Then, around the turn of the first millennium BCE, the first inscription written in Greek (a name written with an inflection characteristic of the Cypriot Greek dialect) is attested, and from there the number of Greek inscriptions increases over time to a boom (in the archaeological record at least) in the 5th and 4th centuries BCE. Although the Greek alphabet had developed and come to be used for Greek elsewhere in the Mediterranean by the 8th-7th centuries BCE, in Cyprus it did not really take hold until the 4th century and beyond, while the Cypriot Syllabary was still being used for Greek in the 3rd or 2nd centuries BCE.

The Cypriot Syllabary is structurally nearly identical to Linear B, and so has very similar deficiencies when writing Greek:

1. It encodes only open syllables: C, CV, and in just two cases, CCV.
2. Its consonant series do not distinguish voicing or aspiration in any cases. The only possible exception is the sign sometimes transcribed as *ga*, although as Egetmeyer has demonstrated at length, there are more complex phonetic issues at play in this case (Egetmeyer 1993; 2010, 188).
3. It does, however, distinguish between /r/ and /l/.
4. It does not distinguish between short and long vowels.

The differences in phonological representation are minimal—the system lacks a separate series for /d/ as found in Linear B, but it does have separate series for /r/ and /l/, unlike Linear B (two issues which may be linked: Steele 2014). The structural mismatch, however, is the same: the signs of the script represent only open syllables while the Cypriot variety of Greek, like all others, has frequent closed syllables in all positions. What is different from Linear B is that Cypriot Syllabic writing involved very different orthographic choices (see Egetmeyer 2010, 220-235):

- Word-final consonants are always written, using a following dummy vowel (in Greek, almost always -se or -ne).
- Consonant clusters are written with plene spelling (i.e., write both consonants but use a “dummy vowel” for the one that has no following vowel).
- Diphthongs are also written with plene spelling (i.e., write one vowel after the other, using the vowel-only sign for the second part).

There is significant standardisation in these spelling strategies, despite a far more varied distribution than we see in Linear B. Within the Mycenaean administrative centres, in documents pertaining to administration, a high level of standardisation is more easily explained by presumed training. But in Cyprus writing was used for a much wider range of purposes: royal inscriptions, administrative documents, religious and funerary monuments, graffiti, and other inscription types, with evidence

for literacy across different social strata and even a small amount of evidence for gender diversity in writing (Steele 2023, 125-137). However, the orthography of Cypriot Syllabic writing presents the writer with fewer choices, since there is a basic underlying rule that every sound is written even if it necessitates the use of a dummy vowel: where Linear B is characterised by both underdifferentiation and underrepresentation of phonemes (Meißner 2007, 97), the Cypriot Syllabary is characterised only by underdifferentiation. At the end of a word, the dummy vowel is always -e, while elsewhere the choice depends on syllabification (typically choosing the vowel of the syllable to which the consonant belongs). This can result in some variation in the spelling of individual words, especially where the syllabification could be interpreted in more than one way: for example, since Greek permits certain consonant clusters at syllable onset, it is sometimes the case that a consonant could be interpreted as belonging to the end of one syllable or the beginning of another, as in the variant spellings *e-se-ta-se* and *e-sa-ta-se* for /estāse/ where the /s/ could close the first syllable or open the second (example from Egetmeyer 2010, 222).

Consani (2016, 101) sees the orthographic rules of the Cypriot Syllabary, which favour the maximum representation of individual phonemes, as a significant factor in the writing system's longevity. Although dummy vowels would have been very frequently encountered, theoretically, it may have taken less cognitive effort to match written sequences with phonological ones, thus aiding word identification and making the writing tradition more accessible to a wider range of people for a wider range of uses. However, it remains that while every sound might be written, not every phoneme was being distinguished. So, for instance, when a Cypriot reader encountered a sequence like *ta-ra-ka-ma-ta*, in order to reach the encoded word /drak^hmata/ they would need to interpret the two instances of *ta* differently (/d/ in one case and /t/ in the other) and the vowel of *ka* as a dummy not to be pronounced. Perhaps in a longer word like this, the overall word shape is helpful in distinguishing between possible words and so making the correct choices for the potentially ambiguous graphemes, while a shorter word could have more potential for confusion. Again it may be worth considering whether there is an economical principle at play, in that differentiating between the full range of phonemes in Cypriot Greek would involve adding multiple extra series of five signs per phoneme to the overall repertoire.

Greek was not the only language written in the Cypriot Syllabary: a number of inscriptions are evidently not written in Greek and can be attributed to at least one other language. The term Eteocypriot has usually been applied to these non-Greek inscriptions, although only a small number exhibit consistent features that appear to belong to the same language (Steele 2013, 99-160). It is possible, perhaps likely, that we are dealing with at least one language that has survived from the Bronze

Age and may once have been written in Cypro-Minoan, although the only feature that clearly seems to link the two is a preponderance of endings in *-ti* in Cypro-Minoan and a common noun ending in *-o-ti* in Eteocypriot. Just as in the case of Linear B we are faced with a question as to how much the structure of the script owes to the language(s) underlying Linear A, for the Cypriot Syllabary we may ask how the structure relates to other languages written in it (or in the earlier Cypro-Minoan). Because a number of Greek names have been borrowed in Eteocypriot, we can see that the orthographic rules seem to have been at least partly shared. However, the spelling of some Greek elements looks unusual, such as examples of both *a-ra-to-* and *a-sa-to-* apparently representing the element *Aristo-* (one of which is in a bilingual inscription where the identification is beyond doubt): are these variant partial spellings of the whole sequence (so potentially evidence of a different orthography), or do they reflect a different phonetic reality that may have resulted from a simplification of a consonant cluster (see Steele 2013, 152)? It is also difficult to discern the features and syllabification patterns of the Eteocypriot language because of its very small corpus (fewer than 30 inscriptions), and so it is difficult to compare its linguistic fit with the Syllabary with that of Cypriot Greek.

The diachronic status of the Cypriot Syllabary looks rather different from that of Linear B. Linear B seems to have changed quite minimally in the couple of centuries of its attested existence, and when it disappeared from the archaeological record it did so suddenly, leaving no traces in the subsequent centuries. It also gives the appearance of disappearing in a linguistic and graphic vacuum, in connection with the changing social and economic landscape, but not obviously in close competition with any other language varieties or writing systems. The Cypriot Syllabary, on the other hand, grew out of Late Bronze Age writing, and was used throughout most of the first millennium BCE, thus with considerably more longevity than Linear B, and in association with a far wider attested range of uses. Literacy in the Cypriot Syllabary, based on surviving evidence, was considerably more widespread than in Linear B. Cypriot Greek speakers seem, at least in some periods, to have thought of themselves as having some affinity with other Greek speakers to the west, perhaps most obvious in historical evidence for the Persian Wars and especially the Ionian Revolt. And yet, Cypriot Greeks did not seem tempted to adopt the Greek alphabet that was so widely used elsewhere, and it was not until certain rulers more consciously began to assert relationships with Athens in the later 5th and 4th centuries BCE that the Greek alphabet came to appear on Cypriot coin legends and occasionally digraphic monuments. By the 4th century, the competing variety of Greek was the pan-Mediterranean Koine, which spread particularly during and after the reign of Alexander the Great and became the official administrative language of Cyprus once the island came under

the jurisdiction of Ptolemy I, after the conflict of the Diadochoi. This top-down change does seem to have impacted wider literate society, but more slowly (see Fig. 1 for a late Cypriot Syllabic religious inscription): the mixed corpus of the religious inscriptions from the 3rd/2nd century BCE Nymphaeum of Kafizin, with both Cypriot Syllabic and alphabetic Koine texts, indicates high levels of contact between the two language varieties and writing traditions; beyond this there is very little trace of the continued existence of the Cypriot Greek dialect or the Cypriot Syllabic writing system (for a fuller account of the loss of this system and introduction of Greek alphabetic writing, see Steele 2018, 219–241).



FIG. 1. Limestone votive relief with a Cypriot Syllabic inscription, 3rd century BCE, from Golgoi. Metropolitan Museum New York. The Cesnola Collection, Purchased by subscription, 1874–76, inv. 74.51.2372

It has often been assumed that the Cypriot Syllabic writing system had some special symbolism or relevance for Cypriot people. For Mitford, it was in the transition to Ptolemaic rule and the disappearance of the local dialect and writing system that ‘Cyprus seemed to lose her soul’ (Mitford 1953, 90). Sherratt has suggested that it was especially the Phoenician expansion and the foundation of the Phoenician colony of Kition on the island in the 9th century BCE that first gave the impetus for Cypriots to embrace the distinctive syllabic writing system as symbolic of their identity (Sherratt 2003; see also Iacovou 2013). But even earlier, in the Late Bronze Age, the visual distinctiveness of Cypro-Minoan went

hand-in-hand with the creative selection of writing practices and supports, which may have looked to both east and west for inspiration (clay document types, seals, etc) but was always idiosyncratically Cypriot (see Steele 2018, 35-44; Steele and Boyes 2023). Whatever the genesis of the link between Cypriot Syllabic writing and Cypriot identity, it seems to have been deeply rooted—and for the larger part of the first millennium BCE it would have been regularly reinforced for Cypriot individuals, from the coins in their pocket to public and private monuments in their visible landscapes, and a whole range of quotidian uses in between.

The visual distinctiveness of the Cypriot Syllabic system is what gives it the potential for a close relationship with linguistic and cultural identity (see also Steele 2018, 149). This writing system was exclusively associated with the island of Cyprus, and left a unique visual footprint wherever Cypriots used it, both in the everyday visible landscapes of Cypriot cities and sanctuaries, and even abroad—for instance at Panhellenic religious sites like Delphi, or among the international graffiti left on ancient monuments in Egypt (on the latter see Pestarino 2025). Theoretically, a Cypriot Greek speaker could have written down their dialect in the Greek alphabet, but this apparently never happened: the occasional Greek alphabetic inscriptions appearing before the end of the 4th century BCE (i.e., before the alphabet came to be used as a central administration's official script) never contain Cypriot dialectal features (most are in the Koine, with some abbreviated forms also). Likewise, we know of no Eteocypriot inscriptions in anything other than the Cypriot Syllabary. We know that Cypriots had wide-ranging contact outside the island throughout the periods when Cypriot Syllabic writing was in use, and so their choice to use their distinctively Cypriot writing system, both at home and abroad, looks like a positive and deliberate one. The link with Cypriot identity created ongoing motivation to maintain the Syllabary, and with it the local language varieties with which it was associated.

We may also ask whether the visual uniqueness of the Cypriot Syllabary was associated with conscious distancing. Were Cypriots specifically concerned with avoiding a linguistic appearance that was not sufficiently visually differentiated from the consonantal alphabets used in the contemporary nearby Levantine coast or the Greek and other alphabets of the wider Mediterranean, or did it only matter to them that the Cypriot language was visually unique? When Cypriot writing was first developed from Linear A in the Bronze Age, it is possible and indeed likely that local and geopolitical identities were of a different nature from those of the later Iron Age: was the choice of the Aegean linear scripts as a model connected with any particular desire to have contemporary Cypriot languages look similar in writing to those of Crete and the Aegean islands, or did it only matter to them that they avoided the appearance of cuneiform associated with the broader Bronze Age Near East?

Finally, it is worth pointing out that there need not be—and in the Cypriot case there was not—a one-for-one match between language variety and writing system. Not only was Cypriot Greek written in the Cypriot Syllabary, but at least one indigenous language was also written in it; the island was also home to a significant Phoenician population using the Phoenician consonantal alphabet. The island's material culture, meanwhile, was quite homogeneous and did not reflect divisions by language (Iacovou 2013, 133). In the age of the independent city kingdoms, individuals who recorded themselves in Cypriot Syllabic writing did not call themselves Cypriots, they referred to themselves as citizens of individual places like Paphos, Marion, or Salamis. On the other hand, names with Kupro- or -kupros elements seem to have been popular for a time in Greek onomastics (e.g., Philokupros). In other words, expressions of group identity were complex and multifaceted, and cannot be reduced to any single association or set of associations. The Cypriot Syllabary could be associated with assertions of linguistic or cultural identity, but should be regarded as just one method of expressing a distinctive identity, available to Cypriot individuals and groups alongside many other visually distinctive aspects of Cypriot culture. Any given individual will have held and presented different aspects of their identity, or different identities, in different ways in different settings.

This section has aimed to use historical examples to demonstrate that a writing system can show traits of vitality (through longevity, usership, domains of use) even when it has a number of apparent deficiencies with regard to its linguistic fit—particularly notable in the case of the Cypriot Syllabary. As we have seen, Cypriots were well aware of an available alternative: the Greek alphabet, which offered much better representation of the phonological inventory of any Greek dialect, a structure that permitted simple notation of consonant clusters without redundancy and an economic principle such that it had considerably fewer graphemes to be memorised—and was used by speakers of other Greek dialects with whom Cypriots acknowledged cultural and political connections (especially obvious in the 5th and 4th centuries BCE). Cypriots did not apparently maintain their syllabic writing system because they could not have improved on it linguistically—they maintained it because it was something undeniably, traditionally, and visibly Cypriot.

3. Visual Distinctiveness in Writing and Its Cultural Context

Unseth argued some time ago that the visual aspect of a writing system is always of high importance to its users, particularly at the point of adoption or creation (Unseth 2005, 2008). He suggested this is especially noticeable from a relational perspective, as users may wish either to assimilate with or distance from other language/writing communi-

ties and to use a visual strategy to do so: in other words, to have their written language explicitly look similar to, or look different from, another written language. This potentially powerful visual principle needs to be contextualised by several sociolinguistic factors, including geopolitics, colonising activity and its legacy, regional and social identities, perceived ethnicities, and so on. It makes a difference who is looking and what the writing landscape looks like in a given place or at a given time—likeness and distinctiveness are relational categories after all. Distinctiveness is also, broadly speaking, a principle that could be applied in several different visual dimensions such as the shapes of graphemes, principles of text arrangement, orthographic strategies and, for want of a better word, “fonts” (i.e., styles of text).

There are thus multiple ways in which a writing system can be visually distinctive, and different possible degrees of visual distinctiveness. For our present purposes I am not concerned with trying to measure visual distinctiveness in any quantitative way, although computational tools could certainly be developed, based on existing models, to do so. Rather, I want to establish the cultural importance of visual distinctiveness in writing, and hence make a case for a visual-cultural category of fit. The groundwork laid in this section will illustrate the complexity of relationships between the parameters identified above (grapheme shapes, text arrangements, orthography, and text styles), now using examples that are in use today rather than historical systems, many of them used to write minority languages.

From the outset, it should be pointed out that not every speaker community will desire to write their language down in a visually distinctive system. For some communities, writing a minority language down in a script used by majority languages could be desirable because it implies equal status or gives a kind of validation of the minority language’s existence. Using a majority script may also bring some advantages for digital literacy. In some areas of the world there is very low diversity of scripts: for instance in Europe there are very few languages written in anything other than the *Latin* or *Cyrillic* alphabets, the most obvious counter example being the longstanding written tradition in Greek—although in some historical periods it would have been less singularly distinctive, particularly in the first half of the first millennium BCE, since the alphabets used for Italic languages, including *Latin*, were much closer in appearance to *Greek alphabets*; they have grown more visually distinctive over time. Today a minority language community like that of *Arvanitika* in Greece can both tell a story about once having had their own distinct alphabet (see <https://hopscotchtranslation.com/2023/02/06/hatzipemou-constantine-arvanitika/>) and at the same time write exclusively in the Greek alphabet of the local majority—and predominantly in Greek language too, as Arvanitika has suffered serious marginalisa-

tion, in part owing to “othering” political narratives about its linguistic relationship with Albanian.

Visual distinctiveness may therefore be highly *unmotivated* for some language communities. However, even if the script is the same, the orthography is quite likely to be noticeably different as the script is used to represent the sounds and structural features of a different language. It is highly unlikely that any two language varieties will look identical when written down—except perhaps in the case of “interscripts,” by which I mean systems that can theoretically use the same orthography for different languages, for example if phonological information is omitted (*Wenyan Classical Chinese* is a good example, or written Putonghua to some extent).

Orthographies of what may be thought of as the same script can be highly visually distinctive. European languages written in the Latin alphabet, for instance, can be recognised quite easily through the distinctive sets of diacritics they use to suit their phonological and orthographic systems. Many orthographies for minority languages employ diacritics in order to mark sounds that do not exist in the writing system used as a base (which is often the Latin alphabet as used for a majority language such as English or Spanish), some requiring a considerable range because of their phonological inventories, including sometimes extra marking for tones. Some also use special characters alongside the borrowed script inventory, like the historical thorn <þ> for /θ/, originally from the Runic alphabet, and edh <ð> for its voiced counterpart /ð/, a historical modification of Latin <d>, still used in the Latin orthography of *Icelandic*.

Because of associations between the Latin alphabet and colonial powers, language communities affected by colonial legacies may seek to disassociate their written language from colonial structures of power or oppression by changing or distancing a Latin orthography. Theoretically, there could be a further advantage that this may also improve the linguistic fit of their writing system, given that many invented Latin orthographies have notoriously underrepresented the phonological inventories of indigenous languages. Spanish orthographies in central America, for instance, are easily recognisable from their distinctive <c> vs <qu> for velar stops footprint, while newer indigenous orthographies tend to undo the orthographic rules surrounding <c> and <q> and make more use of <k>, as seen in spelling reforms such as the *Mayan* language name Cakchiquel > *Kaqchikel*. Investment in distinguishing graphically and visually between Mayan and Spanish languages has even been observed in children (Pellicer 2004). The principle of distancing visually from a colonial language can indeed be so strong that, rather than resulting in better linguistic fit (although this is often outwardly cited as a motivation), such changes can instead make the orthography ‘less phonologically transparent... and less effective in terms of learnability’

(Hull 2017, 142, on *Ch'orti'* Mayan, where for example the connective pronounced [i] is written <yi> purely to distance it from Spanish <y>).

The linguistic fit of an orthography is not necessarily a simple matter of phoneme to grapheme mapping (and hence the expanded categorisations created in Meletis 2020, as we saw above). As the previous example makes clear, the wishes of the speaker community will not be motivated by the same concerns that might occupy linguists designing an orthography. Another often overlooked factor is the compatibility with existing educational practices. For example, Souag has demonstrated that the orthographic strategies of *Ajami Arabic* systems for West African languages, while employing various ways of denoting sounds that do not exist in Arabic, are best understood with regard to their pragmatic roots within Qur'ānic education: in this context it makes sense and is more economical to mark some sounds with a universal “non-Arab sound” diacritic (as opposed to individual phonemic differentiation) (Souag 2010). Although Ajami orthographies are not widely standardised, they are locally well established, and their principles rely on cultural hierarchies of thought, often incorporating locally significant cultural details, that cannot be understood through a European model of how to read a text simply by identifying syntactic linguistic elements (Ngom 2018).

While an orthography can be visually distinctive wherever it is written and in whatever typeface or style, it can potentially also be made even more distinctive through the use of a special font associated strongly with the given language. For instance, *cló Gaelach* (Gaelic type) was used historically for Early Modern *Irish* and continues to be used for Irish and other Celtic languages in the present day—including in socially highly visible contexts such as signage. A similar origin in Early Modern typography can be traced for *Fraktur*, characteristically associated with *German* although now far less frequently used, while the Harri font associated with *Basque*, which has gained in popularity, has its origins in a revival of ancient letter forms. In the *Otomaung* script of Bougainville, one of the three styles of writing employs more ornamental sign shapes that are explicitly said to recall sacred designs of traditional body marking by ceremonial scarification (Kelly 2021, 840-844; Simpson 2024, 68-69).

On the one hand, distinctive styles of writing could be ‘symbolically relevant, acting as distinctive markers of the linguistic landscape... [or as] markers of folklorization or archaization’, triggering mixed reactions among the speaker community (Shah and Brenzinger 2012, 231). While one-on-one visual relationships between language and script style may be desirable to some language communities, for others, they might be seen as restrictive. The domination of a visual style based on early *Cherokee* typography, for example, has more recently been viewed as relatively limiting (in comparison with the vast range of typeface support enjoyed by most majority scripts: see Schillo and Turin 2020 on typeface support for North American minority languages). A recent cam-

paign to change this situation through a crowdfunding initiative (by Typotheque in collaboration with Cherokee type designer Chris Skillern: <https://www.typotheque.com/blog/cherokee-project-update>), is aimed explicitly at ‘preserving Cherokee script diversity and vitality’ (see Fig. 2). It is clear that inequalities in visual representation differ and are felt differently across the spectrum of marginalised—and often digitally disadvantaged—languages.



FIG. 2. ChokeCherry Cherokee Character Set by Chris Skillern. Image courtesy of Chris Skillern

Different parts of the world have very different histories of writing, and attitudes towards particular scripts will at any place, and at any time, depend on a complex nexus of language varieties, writing traditions, group identities, political entities, and many other facets of sociocultural practice and ideology all butting up against each other. For example the wider Balkan area has suffered a notorious history of script appropriation and “othering” against a shifting background of political and cultural upheaval, especially between *Latin* and *Cyrillic* orthographies, with further developments such as the revival of *Glagolitic*: the visuality and visual differentiation of these scripts makes them inevitable targets for political and identarian symbolism (see Selvelli 2021). Depending on the situation and what is available, choices or adoptions of scripts can be

derscoring the relative importance of visual distinctiveness in writing and its links with expressing identity.

It has been noted that a considerable number of recent emergent scripts developed for African languages have followed similar strategies in terms of visual distinctiveness, creating a rich and varied tapestry of writing traditions at different stages of take-up among communities (most recently Brookes 2024, 14-45; Mafundikwa 2004, Tuchscherer 2007, Rovenchak and Glavy 2011 and Unseth 2011 are also helpful and illustrative, although they are quickly becoming outdated in the rapidly shifting landscape of writing throughout the continent). Some of these systems take pre-existing communicative systems as a basis, ones that have typically been excluded from the definition of writing. At this point, it is important to emphasise that there are many marginalised systems around the world that communicate meaning in ways that are not usually recognised (or have not been up to now) in even the most expansive grapholinguistic analyses, from *Nsibidi* and *Adinkra* to *Dongba* and *khipus*. As Martínez-Ruiz has eloquently pointed out, we need to change our viewpoint to understand a system (speaking about what he called graphic writing of the Bakongo in this case) that does not seem to work like writing systems as encountered in the usual taxonomies: “Theories based on linguistic paradigms rooted in Western traditions will not lead to an understanding of Kongo graphic writing systems” (Martínez-Ruiz 2013, 4). Similarly, many systems of the Americas have been overlooked in linguistic taxonomies, despite growing scholarship seeking to address this oversight (e.g., Boone and Mignolo 1994, Salomon and Hyland 2010, Mikulska 2022). It is not that such systems do not encode language (they typically do, including phonography, morphography, and logography to different degrees), but they often do so in ways that are intertwined with other forms of expression and with whole ontologies, philosophies, and worldviews.

An important point arising from the above observation is that, although it is tempting to see it that way, not all systems are necessarily in some kind of (symbiotic or antagonistic) relationship with others nearby, and seeking either to assimilate to or distance from them. A new system created with a more holistic philosophical principle in mind could have cultural benefits because it matches knowledge structures better: after all, a language is itself a system of knowledge. Such knowledge structures may already have forms of visual representation, and developing new writing systems to build on such pre-existing systems of visual representation will have obvious advantages (see Simpson 2024). Resulting emergent systems may indeed seek to “codify” symbol systems to develop something with greater phonographic representation in order to develop wider usability across different media, which may also make them easier to categorise within existing grapholinguistic terms—although in doing so, something may perhaps be lost “in transla-

tion". I will use the example of isiBheqe soHlamvu to illustrate some of the principles that are at stake here.



FIG. 3. Graffiti in isiBheqe soHlamvu. Photograph Courtesy of Sandile Radebe

isiBheqe soHlamvu, also known as Ditema tsa Dinoko, has been developed out of longstanding systems of visual communication to write Ntu languages of southern Africa (see further Auda forthcoming). The prevalence of learning through colonial languages, and corresponding marginalisation of indigenous languages and knowledge systems, has overshadowed education (Mkhize and Ndimande-Hlongwa 2014), and speakers of indigenous languages are impeded in writing their languages down by inconsistent Latin orthographies that vary in grapheme mapping and do not adequately represent their consonants, vowels, or tones. Thus, linguistic fit is a foundational principle for isiBheqe soHlamvu: structurally, it could be called a featural syllabary or abugida, built on triangular base signs with different orientations representing the vowels, to which are added marks (lines, dots, circles, loops, etc) representing consonantal features. The triangle represents the shape of the mouth, meaning that the whole system is built on a visual representation of the points of articulation, with further ways of marking the manner of articulation (e.g., breathy voicing and prenasalisation). Methods of marking tones have also been developed, and the set of *ChiMbire* morphograms created by Shannon Rutendo Goneso (<https://chimbire.africa/>) can be incorporated into writing and read with different phonological realities in different languages. Although L>R linear writing is common, writing

and reading direction are free in isiBheqe, which allows complex layouts (“tessellation” based on the triangular shapes) that are particularly desirable for deliberate aesthetic effects (for instance, in art, crafts, or graffiti; see Fig. 3).

The visual design of isiBheqe is not itself a new creation: it is derived from *izimpawu zesintu* such as the long-established regional symbolologies of *litema tsa basotho*, which communicate knowledge structured through indigenous systems such as the *Nabanotsebo* described by Zulumanthabo Zulu and the concepts and traditions communicated to Western audiences by Credo Mutwa (e.g., Zulu 2014, Mutwa 1965). *Litema tsa basotho* are traditional Sesotho designs used especially in murals on the external walls of houses in Lesotho and South Africa, which communicate information about people’s lives and kinship relations, and consist of geometric patterns in which the triangle is an important element. The ChiMbire morphograms add another dimension to representation since they encode prefixes that are deeply rooted in the epistemological structure of Ntu languages (see Creider 1975, Denny and Creider 1986, Selvik 2001), and so operate on another level of cultural meaning as well as mediating across languages.

Thus, isiBheqe soHlamvu is a script that can potentially encode much more than phonography, depending on the knowledge of the people using it. For those who understand the traditional associations of the designs, it is possible to communicate information related to gender and family associations, for example, with further meaningful use of colour. It is perhaps better to think of the system as encompassing multiple layers, such that it is possible for “lay” writers to simply use the phonographic principles to write (potentially in any language—I have seen transcriptions in Arabic that extend the consonant repertoire to include non-Ntu sounds), and to use colours and arrangements for purely aesthetic effects; while someone who has greater knowledge of the epistemological background of the script, and the visual culture to which it relates, might also communicate more nuanced messages and meanings.

Motivations to create a new writing system vary from one situation to another but there are numerous common threads that have become apparent in communication with minority script users and communities, such as:

- a desire to make a minoritised or indigenous language visible in a landscape visually dominated by majority languages;
- the need to disentangle written language from colonial structures of power;
- principles of visual economy in comparison with existing alphabetic orthographies (the positive assertion that a word can be written with far fewer signs in a new syllabic or semi-syllabic script is surprisingly common);

- the practical aspect of building a writing system from scratch so that it is founded on a principle of accurately representing the features of the target language (and thus a desire for good linguistic fit) and its cultural context.

The last of these points demonstrates how important it is to treat writing system development as a complex concept and process, where visual-cultural (and other aspects of sociocultural) fit and linguistic fit have roles to play to different degrees in different cases. It remains to add that if visual factors can be shown to be of importance, then so too is *visibility*. The case of the Cypriot Syllabary already introduced the idea that high social visibility could correlate with relatively widespread literacy and a wide range of writing domains as markers of high vitality for the writing tradition. How writing looks, and so what language looks like, may be particularly high on the agenda where people are invested in linguistic and cultural identity, for whatever reason.

4. Conclusions

The previous section demonstrated that visual distinctiveness is an important property of writing systems and worth taking into account when we try to explain how a writing system is used and its sociocultural context (or its ‘situational adequacy’: Meletis 2020, 347). Aspects of visual distinctiveness can be described, and there may be some scope in the future to develop ways of measuring it. But can it be diagnostic of the vitality of a writing system? To answer this question we need to consider the possible relationships between visual properties of writing and positive or negative outcomes. The cited historical examples, especially the Cypriot Syllabary, strongly suggest that there is a relationship between visual distinctiveness and indicators of vitality, namely longevity, wide usership and varied domains of usage.

Simpson concludes his survey of visual-cultural features of ten minority scripts by saying that ‘a script’s acceptance ultimately relies on community acceptance. Therefore, language communities contemplating script development could consider implementing visual cultural elements relevant to their unique identities’ (Simpson 2024, 76). Furthermore the incorporation of such elements in the writing system can itself help to preserve traditional culture, making a two-way support relationship between the two. We have seen that writing systems can have strong associations with linguistic identity, because they give a language a unique visual footprint. On the one hand, the design of the writing system may be drawn from existing visual culture. On the other, the visual aspect of the writing may become in some way culturally symbolic: consider the many calligraphic traditions around the world, for example,

where writing is itself art. If “visual-cultural fit” is a useful concept, as I tentatively suggest in this paper, both of these kinds of relationships should be considered as its positive indicators of visual-cultural significance.

Where a writing system has both a good linguistic fit and a good visual-cultural fit, it might be thought to have significant strengths. However, it is important to note that there are other factors involved in vitality, and so there are further matters to address—though these are beyond the scope of the present article. We have already seen that isiBheqe soHlamvu, for example, has crucial advantages both in terms of linguistic and visual-cultural fit. So why is this not a dominant writing system in southern Africa at the time of writing? There are the old problems of finding meaningful state support and putting educational programmes in place to proliferate the system. But digital disadvantage is perhaps the greatest problem in the modern writing landscape: isiBheqe soHlamvu is not yet represented in the Unicode standard at the time of writing, so while there are ways of developing some digital literacy, it remains impossible to make this a reality in the daily lives of most users, where they would want to type it easily on smartphones and computers. A Unicode proposal requires a usage case, and as with many minority scripts, the usage case is hampered by the limitations on digital literacy, creating a vicious circle (see Hossain 2026). Even the existing digital resources for isiBheqe are affected by the legacy of Latin orthographies that have been applied to Ntu languages: for instance, one community-made app (<https://riker.itch.io/isibheqe-translator>) for writing the script only works for isiZulu because it has Latin text input and so requires the orthographic rules that only apply to isiZulu (which has five vowels)—meaning that if you wanted to write in *Sesotho* (which has seven vowels represented by only five graphemes in its Latin orthography) you would often generate the wrong isiBheqe vowel sign and so be unable to write what you intended. The user community will struggle with such issues until better digital support becomes available, just as many other minority writing communities do.

Digital disadvantage is a very significant problem for today’s minoritised, indigenous and endangered languages and their writing systems (Rebin 2024). This does not only apply to newly created and emergent scripts, but also to languages written in orthographies of majority languages—frequent need for complex sets of diacritical marks and sometimes unusual special characters, as designed for fuller phonological representation, means that often they are not sufficiently supported by existing keyboards and typefaces. Indeed, you could say that attempting to achieve full phonological representation, and so a particular type of linguistic fit, has made digital literacy more difficult to accomplish in some cases. The “digital age” offers unprecedented ways of supporting endangered languages, facilitating communication and preserving cul-

tural heritage and knowledge, but the fundamental inequalities in digital access that affect minority communities around the world need to be addressed more effectively to make this a reality (Ajani et al 2024).

Another potential problem for a newly introduced writing system is the possible effect on language learning and literacy, which should also be considered in cases where a visually more distinctive system might be newly introduced. For anyone already literate in another system, there will be a certain learning burden to acquire or switch to the new system; that may be exacerbated when changing from one system to another for the same language. Whether learning a second script that has a very different visual appearance from a learner's first script has a noticeable effect, positive or negative, on acquisition of the new script is a question that requires further research. Gnanadesikan has argued that, for any already literate individual, learning a second script (an S2, like an L2 in multilingual situations) is affected by a sort of cognitive preeminence of their first script: 'learners find them hard to process, use them with an accent, and experience transfer from their S1' (Gnanadesikan 2021, 119). In popular imagination script signs and language sounds tend to be conceptually conflated (as I am often reminded when observing my son's alphabet cartoons), and changing script or orthography entails a significant change to the mental mapping of the language units and so the whole surrounding system of thought.

However, a new system devised with linguistic fit in mind could have particular advantages for child learners, especially if introduced at a time when they are consolidating their acquisition of the language written in it. The same could be said for previously illiterate adult individuals. It must be acknowledged that the situation with regard to language and writing use and literacy will be different in any community, and choices and preferences have to be tailored to individual situations. The relative advantages and disadvantages of a new system could be distributed unevenly across the population. For example, disadvantages related to adult learning could be offset by advantages in child learning, and the latter could be bolstered by beneficial effects of impetus from youth movements and enthusiasm among younger generations. For minoritised languages this is crucial, because low intergenerational transmission—i.e., the failure to pass on the language to the young—is a primary indicator of language endangerment in every scale developed to measure the concept. Any factors that might support a language and its writing system in gaining and maintaining young learners will also bolster their vitality. No system is objectively too new or too complex or too different, it is only within its specific sociocultural setting that its advantages and disadvantages can be weighed against each other.

The three factors cited by Unseth as needed by a writing system to survive (community endorsement, a corpus of literature and a schooling programme: Unseth 2011, 31) are indeed important if not vital to success-

ful take-up, proliferation and maintenance. But how to achieve them is another matter. Community interest and investment are absolutely critical because writing systems need groups of people to put time and effort into teaching, learning and using them. That requires resources (increasingly digital ones alongside more traditional print publications, as well as venues and funding), but it also requires motivation. Motivation is not an abstract concept: it consists of social and personal responses and opinions. Linking a writing system with existing ideas about culture and identity, whatever those may be for a given group of people at a given time, gives it relevance. That is precisely why the visual aspects of a writing tradition matter: social visibility and cultural relevance can impart crucial advantages in the struggle to develop or maintain lasting literacy.

I hope to have achieved three goals through this paper. The first was to show that the category of sociolinguistic fit as offered by Meletis can be productively broken down into further areas of enquiry, rather than bundling together different types of social phenomena as situational context. This allows better comparison and integration with linguistic and other categories of fit, including the visual-cultural category proposed here. The second was to show that visual aspects of writing constitute an important area of study in their own right, in this case by interrogating the principle of visual distinctiveness as a property of writing. The third goal was to show that drilling down on visual properties and visual contexts of writing has real world applicability for minority writing systems, where several recurring factors have been observed in the complex relationships between visual aspects of writing systems and their vitality. Cultivating links with cultural traditions and identities, through visual means, can be key to the motivation needed to maintain a heritage system or propagate an emergent one.

Acknowledgements

This research was conducted as part of the project VIEWS: Visual Interactions in Early Writing Systems, UKRI Frontier Research Grant (Horizon Guarantee) no. EP/X028240/1. I am grateful to (in alphabetical order by first name) Chris Miller, Gerald Roche, Iris Auda, Nick Dunlavy, Olgierd Uziembło, Piers Kelly, Philip Boyes, Stanisław Pstrokoński, Terry Joyce, Tian Tian, Tim Brookes and others for various pieces of advice and feedback that have greatly improved the content of this paper. All remaining errors and omissions are my own. The examples used in the paper are intended to be illustrative rather than exhaustive and reflect some areas of my own research rather than any attempt at even coverage.

References

- Abu Ahmad, H. and D. L. Share (2024). "Beginning reading instruction: Syllables or phonemes? An experimental training study with Arabic-speaking preliterate preschoolers". In: *Developmental Psychology*.
- Ajani, Y. A. et al. (n.d.). "Revitalizing Indigenous Knowledge Systems via Digital Media Technologies for Sustainability of Indigenous Languages". In: *Preservation, Digital Technology & Culture* 53 (), pp. 35–44.
- Auda, I. (forthcoming). "isiBheqe soHlamvu: A Southern African Writing System". In: *Writing As Visual Experience*. Ed. by P.M. Steele. Berlin/Boston: De Gruyter.
- Baroni, A. (2011). "Alphabetic vs. non-alphabetic writing: Linguistic fit and natural tendencies". In: *Rivista di Linguistica* 23.2, pp. 127–159.
- Boone, E. H. and W. D. Mignolo, eds. (1994). *Writing Without Words: Alternative Literacies in Mesoamerica and the Andes*. Durham, NC.
- Brookes, T. (2024). *An Atlas of Endangered Alphabets*. London.
- Cahill, M. (2004). "From Endangered to Less Endangered: Case Histories from Brazil and Papua New Guinea". In: *SIL Electronic Working Papers* 2004-004.
- (2014). "Non-linguistic factors in orthographies". In: *Developing orthographies for unwritten languages*. Ed. by M. Cahill and K. Rice. Dallas, pp. 9–25.
- Consani, C. (2003). *Sillabe e sillabari fra competenza fonologica e pratica scritta*. Alessandria.
- (2016). "In search of the "perfect fit" between speech and writing. The case of the Linear B writing". In: *Variation within and among Writing Systems. Concepts and Methods in the Analysis of Ancient Written Documents*. Ed. by P. Cotticelli-Kurras and A. Rizza. Wiesbaden, pp. 89–104.
- Creider, C. A. (1975). "The Semantic System of Noun Classes in Proto-Bantu". In: *Anthropological Linguistics* 17.3, pp. 127–138.
- Dawson, J. (1989). "Orthography Decisions". In: *Notes on Literacy* 57, pp. 1–13.
- Denny, J. P. and C. A. Creider (1986). "The Semantics of Noun Classes in Proto Bantu". In: *Noun Classes and Categorization: Proceedings of a symposium on categorization and noun classification, Eugene, Oregon, October 1983*. Ed. by C. G. Craig. Vol. 7. Typological Studies in Language, pp. 217–240.
- Diallo, M. S. (2021). "Alphabet ADLaM, au secours de la sauvegarde des langues guinéennes en danger de mort". In: *Editions Francophones Universitaires d'Afrique* 1.3, pp. 403–428.
- Egetmeyer, M. (1993). "Kyprisch za: ein lautliches oder ein graphisches Problem?" In: *Kadmos* 32, pp. 145–155.
- (2010). *Le dialecte grec ancien de Chypre*. Berlin: De Gruyter.
- Fishman, J. A. (1991). *Reversing language shift*. Clevedon: De Gruyter/Brill.

- Frost, R. (2012). "Towards a universal model of reading". In: *Behavioral and Brain Sciences* 35.5, pp. 263–279.
- Gnanadesikan, Amalia E. (2020). "S₁: The Native Script Effect". In: *Proceedings of Grapholinguistics in the 21st Century, 2020*. Ed. by Yannis Haralambous. Vol. 4. Grapholinguistics and Its Applications. Brest: Fluxus Editions, pp. 103–123. DOI: <https://doi.org/10.36824/2020-graf-gnan>.
- Grenoble, L. A. and L. J. Whaley (2005). *Saving Languages: An Introduction to Language Revitalization*. Cambridge University Press.
- Hossain, A. (2026). "Unsettled Scripts: Script Invention and the Bind of Digital Recognition". In: *Modern Languages Open* 1, pp. 1–25.
- Hull, K. (2017). "When letters represent more than sounds: Ideology versus practicality in the development of a standard orthography for Ch'orti' Mayan". In: *Creating Orthographies for Endangered Languages*. Ed. by M. C. Jones and D. Mooney. Cambridge, pp. 142–154.
- Iacovou, M. (2013). "The Cypriot Syllabary as a royal signature: the political context of the syllabic script in the Iron Age". In: *Syllabic Writing on Cyprus and Its Context*. Ed. by P. M. Steele. Cambridge, pp. 133–152.
- Jones, M. C. and D. Mooney, eds. (2017). *Creating Orthographies for Endangered Languages*. Cambridge.
- Joyce, T. and H. Masuda (2019). "On the notions of graphematic representation and orthography from the perspective of the Japanese writing system". In: *Written Language and Literacy* 22.2, pp. 248–280.
- Joyce, T. and D. 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.
- Katz, L. and R. Frost (1992). "The Reading Process is Different for Different Orthographies: The Orthographic Depth Hypothesis". In: *Orthography, Phonology, Morphology, and Meaning*. Ed. by R. Frost and L. Katz. Amsterdam, pp. 67–84.
- Kelly, P. (2018a). "The art of not being legible". In: *Terrain* 70. URL: <http://journals.openedition.org/terrain/17103>.
- (2018b). "The invention, transmission and evolution of writing: Insights from the new scripts of West Africa". In: *Paths into Script Formation in the Ancient Mediterranean*. Ed. by S. Ferrara and M. Valério. Rome, pp. 189–210.
- Kelly, Piers (2020). "The Naasioi Otomaung Alphabet of Bougainville. A Preliminary Sketch From Afar". In: *Proceedings of Grapholinguistics in the 21st Century, 2020*. Ed. by Yannis Haralambous. Vol. 5. Grapholinguistics and Its Applications. Brest: Fluxus Editions, pp. 825–846. DOI: <https://doi.org/10.36824/2020-graf-kell>.
- Lee, N. H. and J. Van Way (2016). "Assessing levels of endangerment in the Catalogue of Endangered Languages (ELCat) using the Language Endangerment Index (LEI)". In: *Language in Society* 45, pp. 271–292.

- Lee, N. H. and J. Van Way (2018). "The language endangerment index". In: *Cataloguing the World's Endangered Languages*. Ed. by L. Campbell, J. Van Way, and A. Belew. Oxford–New York, pp. 66–78.
- Lüpke, F. (2011). "Orthography development". In: *The Cambridge Handbook of Endangered Languages*. Ed. by P. K. Austin and J. Sallabank. Cambridge: Cambridge University Press, pp. 312–336.
- Mafundikwa, S. (2004). *Afrikan alphabets: the story of writing in Africa*. West New York: Mark Batty.
- Martínez-Ruiz, B. (2013). *Kongo Graphic Writing and Other Narratives of the Sign*. Philadelphia: Temple University Press.
- Meißner, T. (2007). "Notes on Mycenaean spelling". In: *Cambridge Classical Journal* 53, pp. 96–111.
- Mikulska, K. (2020). "Introduction". In: *Indigenous Graphic Communication Systems: A Theoretical Approach*. Ed. by K. Mikulska and J. A. Offner. Denver: University Press of Colorado, pp. 3–22.
- (2022). "Homo alphabeticus, la definición de la escritura y las escrituras mesoamericanas". In: *Revista española de antropología americana* 52.2, pp. 201–224.
- Mitford, T. B. (1953). "The character of Ptolemaic rule in Cyprus". In: *Aegyptus* 33.1, pp. 80–90.
- Mkhize, N. and N. Ndimande-Hlongwa (2014). "African Languages, Indigenous Knowledge Systems (IKS), and the Transformation of the Humanities and Social Sciences in Higher Education". In: *Alternation* 21.2, pp. 10–37.
- Morpurgo-Davies, A. (1979). "Terminology of power and terminology of work in Greek and Linear B". In: *Colloquium Mycenaeanum*. Ed. by E. Risch and H. Mühlestein. Neuchâtel-Genève: Université de Neuchâtel, pp. 87–108.
- Mutwa, V. C. (1965). *Indaba, My Children: African Tribal History, Legends, Customs And Religious Beliefs*. London: Canongate Books.
- Ngom, F. (2018). "Ajami literacies of west Africa". In: *Tracing Language Movement in Africa*. Ed. by E. A. Albaugh and K. M. de Luna. Oxford: Oxford University Press, pp. 143–164.
- Pellicer, A. (2004). "The orthographic contrast between two languages: Mayan and Spanish". In: *Written Language and Literacy* 7.1, pp. 35–48.
- Perfetti, C. A. and L. N. Harris (2013). "Universal Reading Processes Are Modulated by Language and Writing System". In: *Language Learning and Development* 9, pp. 296–316.
- Pestarino, B. (2025). "'I see wonderful things!'" Socio-cognitive experience of Cypriot graffiti in ancient Egypt". In: *Journal of Ancient History*, pp. 1–24.
- Rebin, B. (2024). "Paving Future Pathway for Disconnected Voices to Unbalanced Digital World: An analysis of multi-stakeholder perspective on improving the digital support for digitally-disadvantaged languages". MA thesis. Uppsala University.

- Rogers, H. (1995). "Optimal orthographies". In: *Scripts and literacy: Reading and learning to read alphabets, syllabaries and characters*. Ed. by I. Taylor and D. R. Olsen. Dordrecht: Kluwer Academic/Plenum Publishers, pp. 31–43.
- Rovenchak, A. and J. M. Glavy (2011). *African writing systems of the modern age: the Sub-Saharan region*. New Haven: Athinkra.
- Rovenchak, A. et al. (2016). *Proposal for encoding the Mandombe script in the SMP of the UCS (revised)*. URL: <https://www.unicode.org/L2/L2016/16019-mandombe-rev.pdf>.
- Salomon, F. and S. Hyland (2010). "Introduction". In: *Ethnohistory* 57.1, pp. 1–9.
- Sarró, R. (2018). "Between writing and art: The invention of Mandombe". In: *Terrain* 70. URL: <http://journals.openedition.org/terrain/17289>.
- (2023). *Inventing an African Alphabet: Writing, Art, and Kongo Culture in the DRC*. Cambridge: Cambridge University Press.
- Schillo, J. and M. Turin (2020). "Applications and innovations in type-face design for North American Indigenous languages". In: *Book* 2.0 10.1, pp. 71–98.
- Schmidt, D. (2024). "NLP for Readability, Graded Literature, & Materials Development in Tibetan". In: *Revue d'Études Tibétaines* 72, pp. 70–85.
- Seidenberg, M. S. (2011). "Reading in Different Writing Systems: One Architecture, Multiple Solutions". In: *Dyslexia across Languages: Orthography and the Brain-Gene-Behavior Link*. Ed. by P. McCardle et al. Baltimore: Paul H. Brookes Publishing, pp. 146–168.
- Selvelli, G. (2021). *The Alphabet of Discord: The Ideologization of Writing Systems on the Balkans since the Breakup of Multiethnic Empires*. Stuttgart: Ibidem Press.
- Selvik, K. A. (2001). "When a Dance Resembles a Tree. A polysemy analysis of three Setswana noun classes". In: *Polysemy in Cognitive Linguistics: Selected papers from the International Cognitive Linguistics Conference, Amsterdam, 1997*. Ed. by H. Cuyckens and B. E. Zawada. Vol. 177. Current Issues in Linguistic Theory, pp. 161–184.
- Shah, S. and M. Brenzinger (2021). "Writing our language". In: *Revitalizing Endangered Languages: A Practical Guide*. Ed. by J. Olko and J. Sallabank. Cambridge: Cambridge University Press, pp. 220–234.
- Sherratt, E. S. (2003). "Visible writing: questions of script and identity in Early Iron Age Greece and Cyprus". In: *Oxford Journal of Archaeology* 22, pp. 225–242.
- Simpson, L. (2024). "From Icons to Identities: Analysing Visual Cultural Elements in Emerging Scripts". In: *Visible Language* 58.2, pp. 42–81.
- Souag, L. (2010). *Ajami in West Africa*. URL: <https://eprints.soas.ac.uk/13429/>.

- Steele, P. M. (2013). *A Linguistic History of Ancient Cyprus: The Non-Greek Languages, and their Relations with Greek, c. 1600-300 BC*. Cambridge: Cambridge University Press.
- (2018). *Writing and Society in Ancient Cyprus*. Cambridge: Cambridge University Press.
- (2023). *Exploring Writing Systems and Practices in the Bronze Age Aegean*. Oxford: Oxford University Press.
- Steele, P. M. and P. J. Boyes (2023). “A comparative approach to methods of inscribing clay tablets: interaction and innovation in Cyprus and Ugarit”. In: *Current Approaches to the Materiality of Texts in Greco-Roman Antiquity*. Ed. by I. Bultrighini and E. M. Anglicker. Turnhout: Brepols, pp. 35–52.
- Steele, P. M. and T. Meißner (2017). “From Linear B to Linear A: The problem of the backward projection of sound values”. In: *Understanding Relations Between Scripts: The Aegean Writing Systems*. Ed. by P. M. Steele. Oxford: Oxford University Press, pp. 93–110.
- Thompson, R. J. E. (2024). “The Mycenaean Language”. In: *The New Documents in Mycenaean Greek*. Ed. by J. Killen. Cambridge: Cambridge University Press, pp. 232–254.
- Thompson, R. J. E. and T. Meißner (2024). “The Mycenaean Writing System”. In: *The New Documents in Mycenaean Greek*. Ed. by J. Killen. Cambridge: Cambridge University Press, pp. 95–136.
- Tournadre, N. and Gsang-bdag-rdo-rje (2003). *Manual of standard Tibetan: Language and civilization: Introduction to standard Tibetan (spoken and written) followed by an appendix on classical literary Tibetan*. Ithaca, N.Y.: Snow Lion Publications.
- Unseth, P. (2005). “Sociolinguistic parallels between choosing scripts and languages”. In: *Written Language and Literacy* 8.1, pp. 19–42.
- (2008). “The sociolinguistics of script choice: An introduction”. In: *International Journal of the Sociology of Language* 192, pp. 1–4.
- (2011). “Invention of scripts in Africa for ethnic revitalization”. In: *The Handbook of Language and Ethnic Identity. Volume 2: The Success-Failure Continuum in Language and Ethnic Identity Efforts*. Ed. by J. A. Fishman and O. García. Oxford: Oxford University Press, pp. 23–32.
- Zulu, Z. (2014). *The sacred knowledge of the desert: African philosophical transcendence*. Midstream: Madisebo University College Print.

Do We Process Characters as Shapes, Nothing Special?

Mary C. Dyson & David Březina

Abstract. This paper addresses the question of whether we use general object mechanisms to process characters or more specialised processes. A multidisciplinary literature review includes examples of potential specialisations in other domains, such as speech or face recognition, introducing specific forms of expertise. A terminology which clarifies the relationships and representations in the character recognition process is used in identifying similarities and differences between attributes of shape perception and character recognition. The importance of the visual similarities among characters within a font and the retention of font information suggest that expertise in character recognition reflects modifications to our general perceptual processes, i.e., specialisation.

1. Introduction

One debate among psychologists and other scientists is whether we have a few general perceptual mechanisms that we apply to various activities, or more specialised mechanisms that have adapted to particular needs. General mechanisms provide a more parsimonious explanation of how we perceive the world and can be intuitively appealing. In applying this distinction to reading, the question is whether we have specialised mechanisms for character recognition or use general perceptual mechanisms employed in object recognition. By overlooking certain aspects of how we recognise characters when reading, we may be misled into simplifying the perceptual process.

This paper discusses attributes of characters to avoid an oversimplified description of potential character shapes. The focus is on character recognition, describing properties associated with perceptual expertise.

Mary C. Dyson  0000-0002-0920-4312

Department of Typography & Graphic Communication, University of Reading, Reading, RG6 6BZ, UK

E-mail: marycdyson.watkins@gmail.com

David Březina  0009-0007-4410-0133

Rosetta Research

E-mail: hello@mrbrezina.com

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 47–64. <https://doi.org/10.36824/2024-graf-dyso>
ISBN: 978-2-487055-10-0, e-ISBN: 978-2-487055-11-7

The multidisciplinary review aims to demonstrate and clarify the complexity of the process and its implications for graphemics, which may also be of significance for educators in terms of language acquisition. For typographers, there is the practical objective of improving the legibility and readability of text, as well as an interest in elucidating the contribution of fonts to visual perceptual processes. Such knowledge should help eliminate, or at least question, any assumptions by investigators that characters are neutral stimuli. This tendency to treat all characters in the same way, ignoring the differences among fonts, was pointed out by Gilmore (1985) with reference to comparisons of letter confusion matrices.

The premise is that the object-based recognition system is modified to recognise characters as we learn to read. Although our review focuses on behavioural studies rather than neuroscientific research, the “neuronal recycling hypothesis” (Dehaene & Cohen, 2011) provides a helpful description of the process by which the novel function of reading recycles existing brain circuitry. We examine the nature of the perceptual processes that are peculiar to character recognition to identify potential modifications.

1.1. Related Specialisation Debates

In arguing against the special nature of speech perception, Carbonell and Lotto (2014) point out the considerable debate as to whether the accomplishment of speech perception is the result of general perceptual/cognitive mechanisms or special processes. They conclude that “speech sounds are... over-learned categories that play a functional role in a larger hierarchical linguistic system” (p3), but this does not mean that they use different processing mechanisms than other sounds.

In contrast, there has been greater consensus regarding the special status of face perception, distinguishing it from the perception of objects (e.g., Farah et al., 1998). One of the main reasons to attribute a special status to face recognition is the face inversion effect. Most objects or scenes are more difficult to recognise when they are inverted, but faces are considerably more difficult. Thompson (1980) provided a compelling demonstration of the difficulty of perceiving facial expressions in inverted faces. Spatial relations among features that are crucial in the perception of upright features are not apparent when upside down. However, this inversion effect (Yin, 1969) is not unique to face processing and might instead be attributed to expertise (Diamond & Carey, 1986; McGugin & Gauthier, 2010). In the field of design, perceptual processes may also be influenced by expertise. Typographers are more impaired in their judgements of fonts than non-typographers when the characters are inverted relative to upright (Dyson & Stott, 2012).

Both speech and face perception can be considered examples of perceptual over-learning and expertise, with typographic training leading to a more specific expertise. Skilled reading is also a form of expertise.

2. Character and Word Recognition Models

There is broad agreement amongst reading researchers that word recognition is letter-based. Expert readers identify individual letters in parallel. These letters are matched against potential candidate words. In this paper, we focus on character recognition, i.e., the three levels at the bottom of Figure 1.

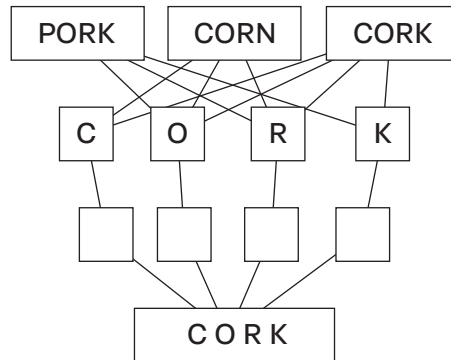


FIG. 1. Word recognition model (after Larson, 2005). The first level represents the stimulus or specific material instance (see below), the next level up is the features used to recognise individual characters (not dealt with in this paper), and the third level is the character detectors. The last, bottom, level represents the identified word.

2.1. Terminology Across Disciplines

Graphematics, psychology, epigraphy, and typography use somewhat disparate terminologies to explore the nature of the relationships and representations involved in the character recognition process. They, naturally, take into account considerations relevant to their respective fields and not necessarily those of the other fields. Graphematics focuses on mapping the units of language and writing systems to their visual representation (see Figure 2). Depending on the source, terms such as “glyph,” “graph,” or “character” can become more or less general and blur the distinction between the general/abstract and material representations that

is highly relevant here. We use a hybrid approach, a terminology of visual communication expanded to account for other fields, in order to be sensitive to the effects of production and perception (see Figure 3).

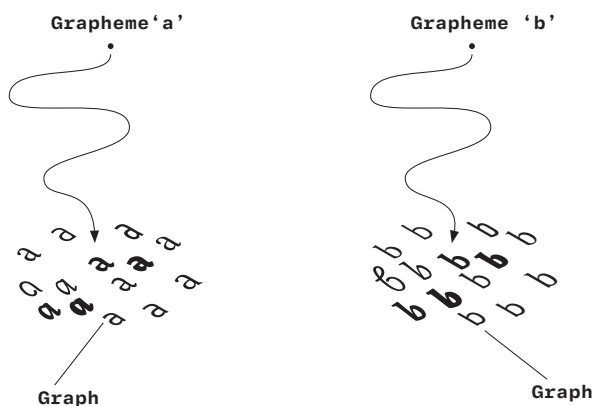


FIG. 2. Linguistic units mapped to images/shapes

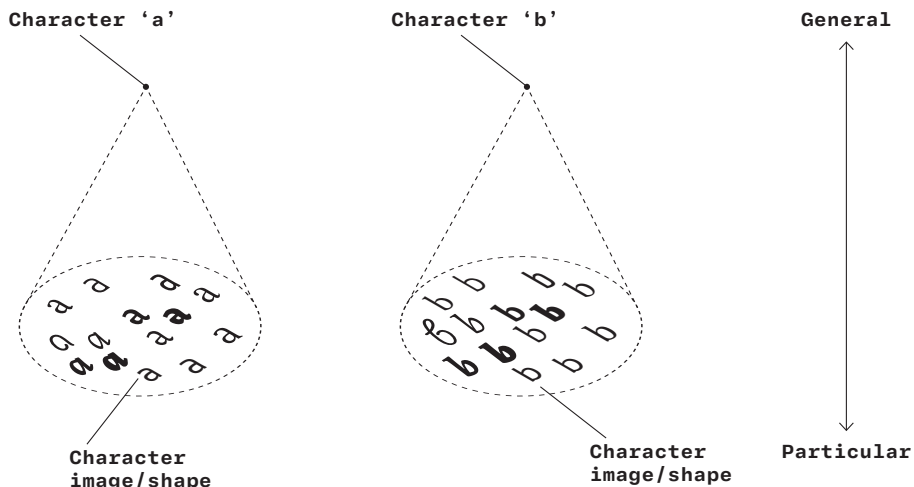


FIG. 3. Hybrid terminology used in this paper clarifies the differences between particular material instances and more abstract character representations.

The term *character image* represents the image on the retina or the specific material instance, while *character shape* refers to the perceptual

representation. The *character* is the most general term referring to the whole class of images/shapes used to represent a particular linguistic unit (comparable to *basic shapes* in Meletis, 2020).

There are class $\rightarrow$ instance/token relationships among some of the terms. With respect to external processes such as production or analytical modelling, it is: character $\rightarrow$ character image. With respect to perception, recognition, and internal processes, the orientation is reversed, from instances to classes: character image $\rightarrow$ character. Characters defined this way are general and visual entities without any inherent linguistic value. One or more characters are associated with one or more linguistic units (e.g., phonemes, syllables, words) from a language by means of a writing system and graphemes (Meletis, 2020).

3. Invariance and Perceptual Constancy

Our perceptual world is more constant than would be predicted by the visual information. An important attribute of object recognition is our ability to perceive an object as having the same shape when viewed at different angles, i.e., shape constancy (Pizlo, 1994). Shape is a visual property of objects that can be used to recognise them.

Readers can cope with many types of visual distortions, e.g., scaling, mirror reflection, rotation, perspective distortions, or environmental effects such as poor lighting or fog, and take account of these distortions during character recognition (see Figure 4). Perceptual constancy describes how we respond to invariants in the physical world.

Readers can also recognise most letters quickly regardless of the font used or style of handwriting (see Figure 5). Consequently, a variety of character images leads to the same character identity. Constancy enables the different character images to be translated from their variant visual forms into invariant representations, referred to as abstract letter identities (Grainger, Rey, & Dufau, 2008).

However, distortions and design variety can lead to uncertainty regarding character identities. In Figure 6, Palmer (1999) illustrated that readers use context to affirm character identities. In the McClelland and Rumelhart (1981) interactive activation model, this is described as top-down processing from the word level interacting with bottom-up processing from the feature and letter level to determine our perceptions.

3.1. From Invariance to Discrimination

Some invariances are problematic for Latin characters, where we should not be applying our normal object recognition processes. Unlike objects, manipulations to visual forms which change orientation, i.e., mir-

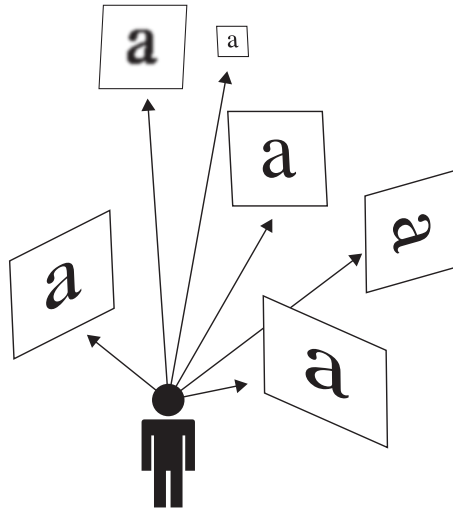


FIG. 4. Distortions caused by different viewing angles and distances

ror imaging (vs. <d>) or plane rotation (<n> vs. <u>) cannot be ignored (Dehaene, 2009; Torres et al., 2021). Although these are not always precise mirror images (see Figure 7), they are sufficiently close to cause problems when the characters need to be discriminated and not perceived as variants of the same form.

From a developmental perspective, Perea, Moret-Tatay, and Panadero (2011) found that by ten years old, children are similar to adult readers in their ability to discriminate between mirror-reversed letters. They propose that learning to read involves “functional tuning of the letter identification system” (p. 244). In the course of literacy acquisition, there is debate as to whether mirror invariance is unlearned, suppressed, or inhibited (Fernandes, Arunkumar, & Huettig, 2021). Unlearning mirror invariance for letters is regarded as a byproduct of neuronal recycling induced by reading (Pegado et al., 2014).

3.2. Is Mirror Discrimination Specific to Latin Readers?

Given the role of reading development in the suppression of mirror generalisation, a pertinent question is whether readers of a script without mirrored characters (e.g., Tamil) develop mirror-image discrimination. By comparing illiterate, Tamil monoliterate, and Tamil-Latin-alphabet bi-literate adults, Fernandes, Arunkumar, and Huettig (2021) concluded that mirror-image discrimination depends on learning to read regardless of script. Tamil monoliterates showed good explicit mirror-image

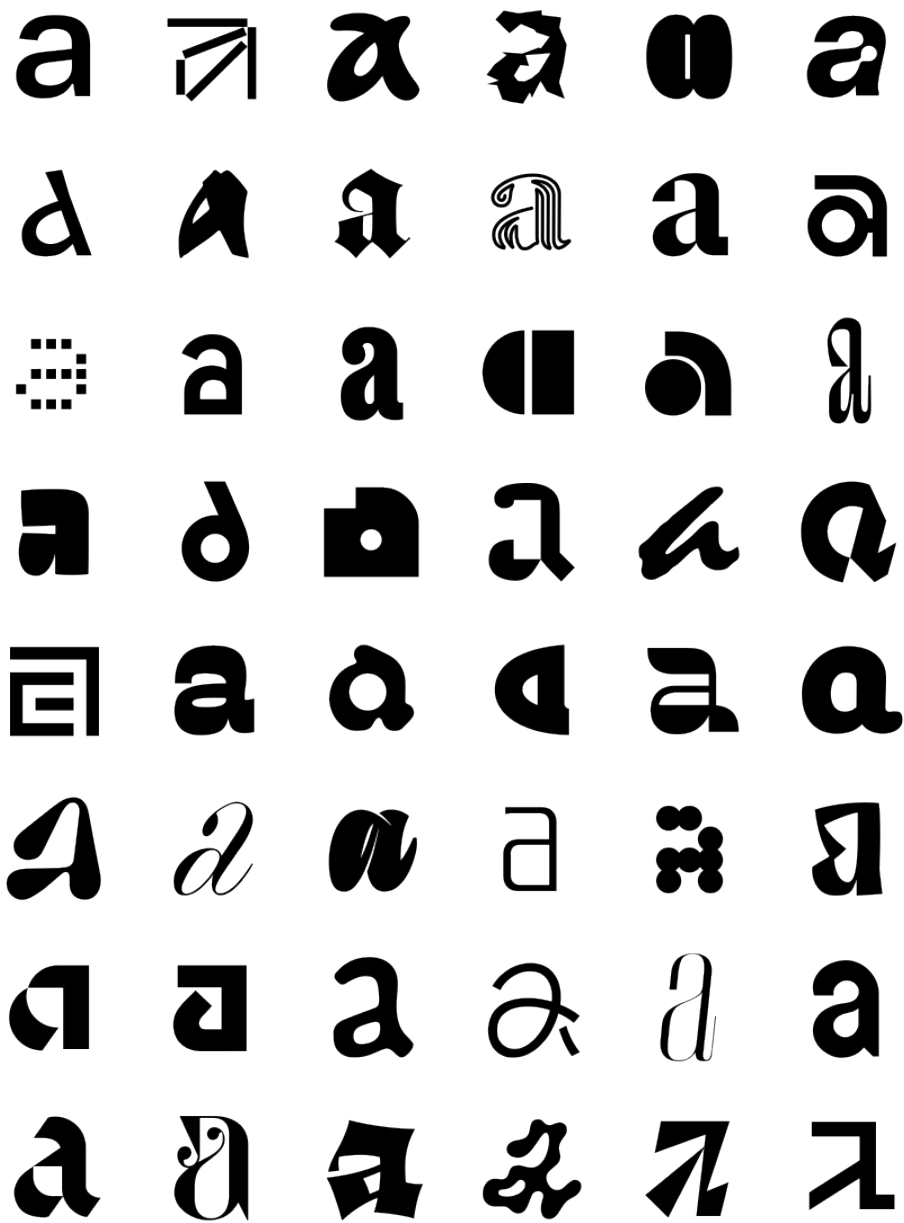


FIG. 5. An inclusive collection of shapes for the letter <a> from various contemporary fonts (Březina, 2024; after Hofstadter & McGraw, 1998)

TAE CAT

FIG. 6. The same shape is used to represent <H> and <A> thanks to its uncertain character identity (after Palmer, 1999). We are able to fluently oscillate between two different character identities thanks to our knowledge of the English language.

bdd pqq



FIG. 7. The letter is not an exact mirror image of the letter <d>. Similarly, the letter <p> is not an exact mirror image of the letter <q>. In both examples, the third letter is the correct form.

discrimination. The authors propose that the ability to discriminate is trained when learning to read because of the specific reading direction. This is supported by a study which found faster responses to Latin characters oriented to the right, when required to judge whether the character was oriented correctly (Patro, Gross, & Friedrich, 2024). The results were consistent across age groups ranging from around seven years to adults.

However, Fernandes, Arunkumar, and Huettig (2021) did find that automatic mirror-discrimination, when not relevant to the task, depends on learning a script with mirrored characters; training from the reading direction is insufficient. They propose that learning a script with mirrored graphs is the strongest force for breaking mirror invariance. Mirror images need to become “graphemically contrastive” (de Heering & Kolinsky, 2019).

4. Visual Coherence

4.1. Type Design

Březina (2018) studied the visual coherence of fonts, i.e., the structure of visual similarities among characters (see Figure 8). He showed that there is a correspondence between the craft practices and the similarity

relationships perceived by readers; that the apparent coherence is created deliberately.

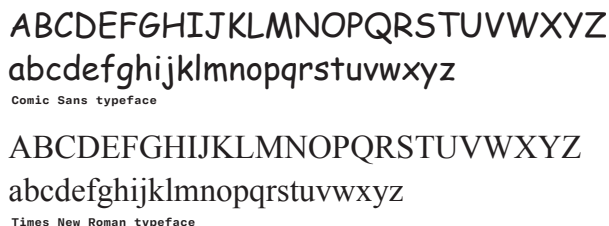


FIG. 8. Visual coherence among character shapes produced by a font makes shapes from the same font look more related to each other than to character shapes produced by a different font (Březina, 2018)

4.2. Character Recognition

We can capitalise on this coherence within a font during character recognition. Expert character recognition takes advantage of the particular characteristics of a font to facilitate recognition, described as a “font-regularity effect” or “font tuning” (Sanocki, 1987). Character processing becomes more efficient because the perceptual system tunes itself to exploit the regularities within a font. The consistency in font styling can be used to translate different character shapes into an invariant (abstract) form. An emphasis on the distinctiveness of characters can overlook the contribution of character uniformity, a view of character processing that considers principles from the field of type design (Sanocki & Dyson, 2012). Although characters may be more distinctive when mixed, they are not easier to read (Figure 9).

**If you don't continue reading the rest of
this paper, you will miss finding out more
about why characters might be special.**

FIG. 9. An extreme example of mixing fonts where we need to adjust the frame of reference for alternating characters.

The original research conducted by Sanocki (1988) used the crude technology of the time to explore font tuning (Figure 10). Subsequently, researchers pursued other enquiries: testing the speculation

that font tuning distinguishes character recognition from object recognition (Gauthier et al., 2006); looking at the time course of font tuning (Walker, 2008); using font tuning as a means of exploring the effectiveness of different type variants in emphasising text (Dyson & Beier, 2016).

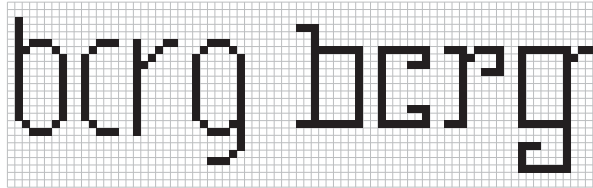


FIG. 10. Fonts used by Sanocki

These studies have employed very different tasks, but all involved brief presentations of characters or words, and readers were required to respond as quickly and accurately as possible. This time pressure is necessary as character recognition is very rapid. Gauthier et al. (2006) make explicit that the artificial tasks are “designed to reduce the impact of the knowledge of orthographic and phonological regularities that develop with reading experience.” (p543).

Sanocki (1988) displayed four character images very briefly, in mixed or consistent fonts. Characters were identified more accurately when the font was consistent.

Many years later, Gauthier et al. (2006) used a search of a matrix where characters were displayed in mixed fonts or all in the same font. Mixing decreased accuracy and increased the time to locate target characters. Evidence for the process of font tuning is mainly based on Latin, but the researchers conducted additional experiments with Chinese characters. They questioned whether experience with characters (reading expertise) is necessary to tune to the stylistic regularities within a font by comparing Chinese and non-Chinese readers. This task required the identification of three characters, presented briefly on a screen. The novice readers were trained with the verbal labels for the Chinese characters before the experiment. They found that non-Chinese readers do not tune to stylistic regularities. Readers who are not familiar with the script appear unable to take advantage of the font regularity and perceive meaningless shapes, rather than characters. They conclude that characters are not processed in the same manner as shapes.

Walker (2008) elaborated upon the concept and nature of font tuning, describing the process as deriving and implementing font-specific rules which map visual features (style-specific) onto structural descriptions (font-invariant). His experiments revealed that we may have some control over the time course of font tuning. In these studies, a random string

of consonants was displayed above or below a word or pseudoword. The task was to identify whether there was a word or not and to ignore the string of consonants. When the consonants were in a (very) different font from the target, responses were slower. Varying the frequency of font repetition across trials appears to influence whether the font-specific translation rules are retained in working memory for reimplementing with the next portion of text.

The study of font tuning by Dyson and Beier (2016) aimed to determine the type of stylistic variation and how much variation is tolerated before character recognition becomes more difficult. This knowledge has practical relevance in text design, where italic or bold type distinguishes a word from the surrounding text. The variation among six fonts was controlled as these were created by Sofie Beier, a type designer. Rapid Serial Visual Presentation of a string of 15 words, with a possible pseudoword inserted, was used to assess the speed and accuracy of detecting the pseudoword. Alternating between a regular and italic font did not show a font-regularity effect, i.e., responses were not slower than a single font. However, alternating regular with a heavier weight, bold, did show this effect. The practical implications for designers are that words in italic can distinguish text elements without significantly disrupting reading. Bold is likely to be more appropriate than italic for headings by making words stand out.

4.3. Objects

Walker (2008) concludes his paper by surmising as to whether object recognition might also be facilitated by the equivalent of font tuning. He uses the example of sets of cutlery in the same design (a similar example appears in Figure 11). Is it easier to identify items within the same set, rather than mixed? To our knowledge, whether stylistic features aid the recognition of objects has not yet been tested.



FIG. 11. Two sets of cutlery (after Walker, 2008): Liner (left) and Provençal (right) from <https://www.davidmellordesign.com/cutlery>

5. Modelling Object vs. Character Recognition

The study described above by Gauthier et al. (2006) concluded that character recognition can be distinguished from object recognition based on font regularity. An alternative means of addressing the distinction between domain-specific (character) and domain-general (object) features concludes that object-based features account well for character perception (Janini et al., 2022). This study used machine learning to model features and focused on the set of features that determine distinctiveness, looking for a match with behavioural data that measured perceptual similarity between characters.

5.1. Perceptual Similarity

In using two behavioural tasks to measure similarity, visual search and letter categorisation, Janini et al. (2022) are assuming that similarity is the inverse of distinctiveness and that this is sufficient to explain the accuracy of character recognition. The discussion of font tuning based on the coherence of characters within a font questions this assumption. Furthermore, Mueller and Weidemann (2012), through a review of legibility literature, experiments, and modelling, refer to the term “perceivability” which they equate with “legibility,” concluding that perceivability should be considered as independent from similarity. Machine learning does not speak directly to which features readers use to recognise characters when reading. This limitation is acknowledged by Janini et al. (2022, p. 13): “Future studies are required to determine the extent to which our findings generalize to the context of reading”.

6. Characters in the Context of Reading

The above studies explored or modelled character recognition through studies of single characters or words, focusing on the early stages of word recognition, primarily the bottom-up processes. This emphasis disregards any potential residual effects of font style information beyond the creation of abstract letter identities. Although there is consensus on the inclusion of abstract letter identities in character identification, this does not preclude the existence of other forms of representations (Walker & Hinkley, 2003).

There is some evidence that font style information can linger beyond the creation of abstract letter representations. In the first experiment conducted by Lewis and Walker (1989) they examined whether the visual style of the characters in a word would be encoded unintentionally and whether this might affect semantic judgements. This was con-

firmed as reaction times to words were slower if the perceptual qualities of the font were inconsistent with the meaning of the word, e.g., the word “heavy” in a “light” font (Figure 12). Their second study widened the scope to include animal words where participants were required to indicate whether the animal was fast/slow or light/heavy as quickly as possible (see Figure 13). These effects suggest that there are some top-down influences on character and word recognition, included in interactive models of reading (McClelland & Rumelhart, 1981).

Consistent	heavy	<i>light</i>
Inconsistent	<i>heavy</i>	light

FIG. 12. Responses were slower if the perceptual qualities of the fonts were inconsistent with the meaning of the words (Lewis & Walker, 1989)

	Fast	Slow	Light	Heavy
Consistent	<i>fox</i>	cow	<i>rat</i>	pig
	<i>deer</i>	bear	<i>mouse</i>	rhino
Inconsistent	fox	<i>cow</i>	rat	<i>pig</i>
	deer	<i>bear</i>	mouse	<i>rhino</i>

FIG. 13. Participants were slower to identify if the animal was fast/slow or light/heavy if the font was inconsistent with this quality (Lewis & Walker, 1989)

This type of inconsistency or incongruence is based on the Stroop task (Figure 14), where we are required to name the colour of each word.

However, the Lewis and Walker implementation is a reverse Stroop effect, as you need to ignore the font and respond to the word's meaning.

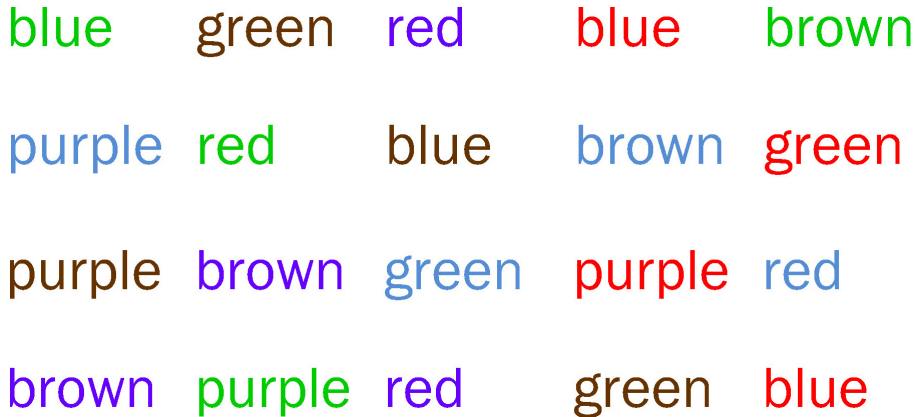


FIG. 14. The original Stroop effect, where participants are slower to name the colour of the word when the meaning is inconsistent with the colour of the ink.

A study concerned with memory for the visual appearance of words found evidence of the “long-term retention of superficial details (i.e., typography)” (Sheridan & Reingold, 2012, p. 255). The study included measurements of eye movements while reading sentences with a low-frequency target word (e.g., *banjo*) in “distorted typographies”. Fixation times for the second presentation of the target word (in a different sentence) were shorter when the visual style was consistent with the first presentation. This effect was found within the relatively short time span of an experiment and with a gap of one week. However, difficult-to-read or unusual fonts were required to demonstrate the effect after a one-week lag. The apparent memory advantage for less legible fonts, which take longer to read, is an example of the reported effects of perceptual disfluency (see Dyson, 2022, for a review).

In a further study concerned with memory for the visual appearance of words, there was some indication that font styling is retained in memory (Dyson & Březina, 2022). Changing the font of an item between an initial lexical decision task (word or non-word) and subsequent presentation impaired recognition. This behavioural data is supported by an event-related potential study showing a response in the brain to the font at a later stage, beyond the assumed point of abstracting letter identity (Keage et al., 2014).

The possibility of stylistic features of characters being retained beyond character recognition and interacting with semantic information would seem to distinguish characters from mere shapes.

7. Conclusion

Table 1 summarises attributes of shapes and characters, identifying similarities and differences. The differences indicate processes that may have been introduced during neuronal recycling, while the similarities indicate processes that appear to be shared. The review highlights that readers retain relevant semantic information from characters outside their direct linguistic function. Stylistic features are used to facilitate recognition across characters from different linguistic categories. The perceptual constancy inherent in object/shape recognition is useful in character recognition when mapping the variant forms onto a single representation. However, where character recognition requires discrimination between different orientations, learning to read introduces modifications to the treatment of invariance. Skilled readers develop a perceptual expertise that reflects the demands of the task.

In graphetics, the world’s scripts are usually viewed as character inventories. The mutual visual relationships of characters within scripts are often described informally. Exploration of the component perceptual processes seems to be a good point of departure for a deeper formal analysis of character relationships and their regularity.

TABLE 1. Similarities and differences between some of the perceptual processes involved in shape and character recognition, indicating potential effects of neuronal recycling

	Shapes	Characters
Size, rotation, position invariance	+	+
Invariance to visual style	+	+
“Unlearned” mirror invariance	x	Latin script
Capitalizes on coherence	?	+
Secondary semantic information	–	+

References

Březina, David (2018). “Coherence in typeface design: visual similarity of characters in Cyrillic, Devanagari, and Latin”. PhD thesis. University of Reading.

Carbonell, K. M. and A. J. Lotto (2014). “Speech is not special... again”. In: *Frontiers in Psychology* 5, p. 427.

Dehaene, Stanislas (2009). *Reading in the brain: the new science of how we read*. Penguin Books.

- Dehaene, Stanislas and Laurent Cohen (2011). "The unique role of the visual word form area in reading". In: *Trends in Cognitive Sciences* 15.6, pp. 254–262.
- Diamond, Rhea and Susan Carey (1986). "Why faces are and are not special—an effect of expertise". In: *Journal of Experimental Psychology: General* 115.2, pp. 107–117.
- Dyson, Mary and David Březina (2022). *The sequel to exploring disfluency: Do we remember the visual appearance of words?* Design Regression. URL: <https://designregression.com/research/the-sequel-to-exploring-disfluency-do-we-remember-the-visual-appearance-of-words>.
- Dyson, Mary C. (2022). "Perceptual Disfluency Through Hard-to-Read Fonts. Is There a Satisfactory Explanation?" In: *Proceedings of Grapholinguistics in the 21st Century, 2022*. Ed. by Yannis Haralambous. Vol. 9. Grapholinguistics and Its Applications. Brest: Fluxus Editions, pp. 101–109. DOI: <https://doi.org/10.36824/2022-graf-dyso>.
- Dyson, Mary C. and Sofie Beier (2016). "Investigating typographic differentiation: italics are more subtle than bold for emphasis". In: *Information Design Journal* 22.1, pp. 3–18.
- Dyson, Mary C. and C. A. Stott (2012). "Characterizing typographic expertise: Do we process typefaces like faces?" In: *Visual Cognition* 20.9, pp. 1082–1094.
- Farah, M. J. et al. (1998). "What is "special" about face perception?" In: *Psychological Review* 105.3, pp. 482–498.
- Fernandes, T., M. Arunkumar, and F. Huettig (2021). "The role of the written script in shaping mirror-image discrimination: Evidence from illiterate, Tamil literate, and Tamil-Latin-alphabet bi-literate adults". In: *Cognition* 206, p. 104493.
- Gauthier, I. et al. (2006). "Font tuning associated with expertise in letter perception". In: *Perception* 35, pp. 541–559.
- Gilmore, G. C. (1985). "Letters Are Visual-Stimuli—a Comment on the Use of Confusion Matrices". In: *Perception & Psychophysics* 37.4, pp. 389–390.
- Grainger, Jonathan, Arnaud Rey, and Stéphane Dufau (2008). "Letter perception: from pixels to pandemonium". In: *Trends in Cognitive Sciences* 12.10, pp. 381–387.
- Heering, A. de and R. Kolinsky (2019). "Braille readers break mirror invariance for both visual Braille and Latin letters". In: *Cognition* 189, pp. 55–59.
- Hofstadter, Douglas R. and Gary McGraw (1998). "Letter Spirit: Esthetic Perception and Creative Play in the Rich Microcosm of the Roman Alphabet". In: *Fluid Concepts & Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. Ed. by Douglas Richard Hofstadter and The Fluid Analogies Research Group. Basic Books, pp. 407–466.
- Janini, D. et al. (2022). "General object-based features account for letter perception". In: *PLoS Comput Biol* 18.9, e1010522.

- Keage, H. A. D. et al. (2014). "Investigating letter recognition in the brain by varying typeface: An event-related potential study". In: *Brain and Cognition* 88, pp. 83–89.
- Larson, Kevin (2005). "The science of word recognition". In: *Typo* 13, pp. 2–11.
- Lewis, C. and P. Walker (1989). "Typographic influences on reading". In: *British Journal of Psychology* 80.2, pp. 241–257.
- 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.
- McGugin, R. W. and I. Gauthier (2010). "Perceptual expertise with objects predicts another hallmark of face perception". In: *Journal of Vision* 10.4, pp. 1–12.
- Meletis, Dimitrios (2020). *The Nature of Writing. A Theory of Grapholinguistics*. Vol. 3. Grapholinguistics and Its Applications. Brest: Fluxus Editions. DOI: <https://doi.org/10.36824/2020-meletis>.
- Mueller, S. T. and C. T. Weidemann (2012). "Alphabetic letter identification: Effects of perceivability, similarity, and bias". In: *Acta Psychologica* 139.1, pp. 19–37.
- Palmer, Stephen E. (1999). *Vision Science: Photons to Phenomenology*. MIT Press.
- Patro, K., A. Gross, and C. Friedrich (2025). "Spatial biases in processing mirror letters by literate subjects". In: *Reading and Writing* 38, pp. 821–841.
- Pegado, F. et al. (2014). "Literacy breaks mirror invariance for visual stimuli: a behavioral study with adult illiterates". In: *Journal of Experimental Psychology: General* 143.2, pp. 887–894.
- Perea, M., C. Moret-Tatay, and V. Panadero (2011). "Suppression of mirror generalization for reversible letters: Evidence from masked priming". In: *Journal of Memory and Language* 65.3, pp. 237–246.
- Pizlo, Z. (1994). "A theory of shape constancy based on perspective invariants". In: *Vision Research* 34.12, pp. 1637–1658.
- Sanocki, Thomas (1987). "Visual knowledge underlying letter perception: font-specific schematic tuning". In: *Journal of Experimental Psychology: Human Perception and Performance* 13.2, pp. 267–278.
- (1988). "Font regularity constraints on the process of letter recognition". In: *Journal of Experimental Psychology: Human Perception and Performance* 14.3, pp. 472–480.
- 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.
- Sheridan, Heather and Eyal M. Reingold (2012). "Perceptual specificity effects in rereading: Evidence from eye movements". In: *Journal of Memory and Language* 67.2, pp. 255–269.

- Thompson, Peter (1980). "Margaret Thatcher: a new illusion". In: *Perception* 9, pp. 483–484.
- Torres, A. R. et al. (2021). "Selective inhibition of mirror invariance for letters consolidated by sleep doubles reading fluency". In: *Current Biology* 31.4, 742–752.e748.
- Walker, Peter (2008). "Font tuning: A review and new experimental evidence". In: *Visual Cognition* 16.8, pp. 1022–1058.
- Walker, Peter and L. Hinkley (2003). "Visual memory for shape-colour conjunctions utilizes structural descriptions of letter shape". In: *Visual Cognition* 10.8, pp. 987–1000.
- Yin, R. K. (1969). "Looking at upside-down faces". In: *Journal of Experimental Psychology* 81.1, pp. 141–145.

On the Linguistic Status of Emojis

Dennis Reisloh & Tatjana Scheffler

Abstract. This study explores the linguistic potential of emojis, focusing on their grammatical and semiotic qualities. While previous research has examined emojis as language accompaniments, this paper investigates their use in standalone sequences. In two experiments, participants transferred simple German sentences into emoji sequences, to enable analyses into how a) subject-predicate order and b) human versus animal agents affect emoji usage. The results indicate that although emojis can represent simple sentences while assuming basal grammatical features, a systematic use of grammatical concepts cannot be found. Instead, participants often rely on visual characteristics of the emojis rather than abstract linguistic meaning.

1. Introduction

The increase of digital messages deviating from standard linguistic norms has created a necessity for the expansion of graphic sign systems in digital linguistic spaces (Androutsopoulos, 2020). In these medially written, but mostly conceptually oral (cf. Koch and Oesterreicher, 1985), contexts, emojis are used frequently (Androutsopoulos, 2003; Siebenhaar, 2018). While research has examined the language accompanying use of emojis, identifying their linguistic functions (e.g. Dürscheid and Siever, 2017; Pappert, 2017), fewer studies focus on the use of emojis as standalone sequences. This could be due to the fact that emojis are context-sensitive to a large extent (Gawne and McCulloch, 2019). However, recent work by (Weissman, Engelen, Baas, and Cohn, 2023) shows that emojis are undergoing a process of lexicalization, which allows some of them, e.g., object and activity emojis, to be interpreted independently of context. Furthermore, this raises the question of the extent to which emojis have the potential to be used as language-independent

Dennis Reisloh  0009-0006-5187-7947
Ruhr-University Bochum, Germany
Tatjana Scheffler  0000-0001-7498-6202
Ruhr-University Bochum, Germany

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 65–93. <https://doi.org/10.36824/2024-graf-reis>
ISBN: 978-2-487055-10-0, e-ISBN: 978-2-487055-11-7

linguistic signals. While emojis have been shown to assume graphematic functions (Dürscheid and Meletis, 2019) and substitute for words or punctuation marks (Scheffler, Brandt, Fuente, and Nenchev, 2022), they are missing complex grammatical structures (Dürscheid and Siever, 2017). For these reasons, emojis could not be classified as independent linguistic elements but rather as elements that accompany language. Nevertheless, the pure use of emoji sequences can still be observed (McCulloch and Gawne, 2018). Building on the framework of Jackendoff and Wittenberg (2014), Cohn, Engelen, and Schilperoord (2019) has established a hierarchy for different grammatical levels of emoji sequences, which shows different possible basic grammatical constructions.

This paper aims to expand on current research, exploring the grammatical and semiotic qualities of object and activity emojis in more detail. For this purpose, we devised an experiment that asks participants to translate standard German sentences into emojis. This allows us to answer the following two research questions, providing deeper insight into the linguistic status of emojis:

1. To what extent does the position of the subject in the given sentence affect the position of the corresponding emoji in the translated sequence?
2. Are emojis depicting people from the subcategory “activities and sports” interpreted in a human context and preferentially selected for this purpose or are they used independently to describe an abstract activity?

Both questions address the emojis’ potential as abstract linguistic signals. The first investigates the syntax of emoji sequences (following *ibid.*), while the second investigates whether activity emojis are still interpreted holistically as depicting scenarios, or whether their meaning has conventionalized and become more detached from a specific scene and actor type.

2. Background

Although the grammar of emojis remains a rarely explored topic in current research, a few studies have examined it from different perspectives. Studies have shown that emojis can be substituted for words or punctuation marks (Cohn, Roijackers, Schaap, and Engelen, 2018; Pappert, 2017; Scheffler, Brandt, Fuente, and Nenchev, 2022). McCulloch and Gawne (2018) argue that emojis are most comparable to words; however, this alone is insufficient to classify them as an independent language, as languages must exhibit grammatical structures. They reject the notion that emojis possess a hierarchical grammar, supporting their claim by analyzing the ten most frequently used bigrams, trigrams, and quadrigrams of emoji sequences. With only one exception, these sequences

predominantly consist of repeated instances of the same emoji. Through corpus analyses, they demonstrate that this pattern contrasts with the English language—even in informal contexts—and argue that emojis do not function with an autonomous grammar. Instead, they behave similarly to rhythmic gestures in relation to written text.

This raises the question of how emojis operate when they are not embedded within written text. Cohn, Engelen, and Schilperoord (2019) have categorized different types of emoji sequences and classified them into four distinct grammatical levels, drawing on the hierarchy proposed by Jackendoff and Wittenberg (2014). This hierarchy not only encompasses complex linguistic systems but also differentiates between possible precursors to fully developed language structures.

The lowest category in this classification consists of *one-unit grammars*, which include sequences of multiple emojis used to create a larger unit or a cohesive image. Following this is the category of *linear grammars*, where emojis are employed as lists of different elements or temporal sequences. The frequent *reduplications* of emojis highlighted by McCulloch and Gawne (2018) are also included in this category. Such repetitions function as intensifiers or as indicators of large quantities, analogous to the repetition of morphemes in various languages (Cohn, Engelen, and Schilperoord, 2019).

Although the model excludes more complex grammatical structures for emojis, the next two categories at least exhibit basic forms of grammar. The *categorical grammars* type includes *two-unit* and *three-unit* emoji sequences, which incorporate grammatical elements such as subject, verb, and object. Notably, in the given examples, the sequence of emojis does not necessarily align with the word order of the corresponding translated sentence. Furthermore, the subject of the sentence is not represented by an emoji but is instead assumed to have an automatic first-person indexicality—an observation frequently made in emoji usage (Grosz, Kaiser, and Pierini, 2021).

As the fourth and highest grammatical level for emojis, Cohn, Engelen, and Schilperoord (2019) introduce *simple phrase grammars*. This category, consisting of only one subcategory, is limited to cases where emoji sequences emerge through the embedding of grouped emojis within other grouped emojis. For instance, the role of the subject may not be conveyed by a single emoji but rather by an *affixation*, i.e., a combination of multiple emojis that together form a larger unit. Beyond the grammatical hierarchy, additional phenomena, such as the use of metonymy and rebus writing, are noted. These findings align with observations made by Weissman, Engelen, Baas, and Cohn (2023) on the presence of polysemy and homonymy in emoji usage.

The research conducted by Weissman, Engelen, Baas, and Cohn (ibid.) and Cohn, Engelen, and Schilperoord (2019) serves as a foundation for the first experiment in our empirical study presented here.

It is worth noting that the German language, in which the experiments were conducted, is particularly well-suited for investigating the syntactic structures of emoji sequences. Unlike English, for example, German has a relatively free word order and is not strictly bound to an S-V-O structure. The verb-second (V2) constraint allows for more flexible sentence constructions, including the possibility of placing the subject at the end of the sentence (O-V-S). This flexibility enables a more precise discussion of the results regarding the extent to which participants adhered to the given syntax of the matrix language, German, or modified it in their emoji sequences.

The second experiment focuses on activity-related emojis. Grosz, Kaiser, and Pierini (2021) have already found that, unlike facial emojis, activity emojis are less influenced by first-person indexicality and instead function as event-denoting free adjuncts that assign an individual as the agent of an action. Our experiment builds on this finding by examining two distinct categories of activity-related emojis: those that depict humans performing an activity and those that represent an object symbolically associated with an activity.

By manipulating predefined sentences in which a human agent is replaced by an animal, we aim to determine whether these two categories of activity-related emojis are used differently. Specifically, the analysis will explore whether human-representing emojis are consistently chosen to signify a human agent performing the action or whether the selection is guided solely by the activity itself, independent of the agent's identity.

3. Empirical Study

Former studies have already made evident that complex matters or grammatical structures cannot be adequately represented through emojis (Cohn, Engelen, and Schilperoord, 2019; Dürscheid and Siever, 2017). Following Chomsky's assumption (1985, p. 4) that grammar serves as a description of the idealized competence of language, it is conceivable that patterns in the use of emojis could emerge, even if only basic grammatical properties are present. Consequently, the following study tests a "translation" of simple sentences into freely chosen emoji sequences by participants.

3.1. Methodology

3.1.1. *Participants*

The study was conducted with 49 people via an online questionnaire. We recruited 49 individuals. No data were collected that would allow direct identification of participants. Of the participants, 46 reported German

as their native language, while three stated they were fluent in German. Regarding gender identity, 31 identified as female, 16 as male, and two as another gender. One participant indicated being under 20 years old, while the age ranges of 20–29 and 30–39 were the most represented with 13 and 17 participants respectively. Additionally, five participants reported being between 40–49 years old, nine between 50–59 years old, and four were either 60 years or older. The highest levels of education achieved were varied: one participant is a (secondary) student, five had completed secondary school (“*Realschulabschluss*”), twelve had completed high school (“*Abitur*”), 17 held a vocational qualification, six had a Bachelor’s degree, and eight had earned a Master’s degree.

In order to estimate the participants’ emoji fluency, we asked them to specify how often they typically use emojis in contexts such as chats or social media on a scale from 1 to 5 (1 = never, 2 = rarely, 3 = occasionally, 4 = frequently, 5 = always). The mean score was 4.429 (sd = 0.707), indicating a high frequency of emojis. No participants specified their emoji usage as “never” or “rarely”. When asked about the frequency of exclusively using emojis without additional text in messages or posts, the mean score was 2.694 (sd = 0.822). A detailed overview of the demographic data, presented as bar charts, can be found in Appendix A.

3.1.2. Study Design

Participants were randomly assigned to one of two questionnaires on the online survey platform SoSci Survey¹. The link was programmed to randomly assign one of the two questionnaires upon the first access and to ensure that the next participant received the alternate questionnaire. This process is repeated iteratively. Participants were not informed about their random assignment to one of the two groups. Full versions of the questionnaires are available in the appendix of this study. After providing participants with information about the approximate duration of the survey, a recommendation to use a mobile device (due to its built-in emoji keyboard), and collecting the demographic data summarized in 3.1.1, the task was presented as follows (translated into English): “On the following pages, you will see a short sentence displayed on each page. Please use the input field below to represent the sentence using emojis. The number of emojis you use can be greater or smaller than the number of words in the sentence. Create an emoji sequence that, in your opinion, provides the best possible ‘translation.’ Please use only emojis available on the emoji keyboard of your device and do not use any other symbols.”

Subsequently, participants were presented with ten simple sentences per experiment. Each sentence included an open text field for the partic-

1. <https://www.soscisurvey.de/>

ipants to input their emoji sequences. The 20 sentences used across the two experiments contained a maximum of three sentence constituents and primarily included words that could be represented with emojis from categories known to be easily understood, even out of context (cf. Weissman, Engelen, Baas, and Cohn, 2023). The sentences from both experiments were randomly mixed and displayed individually on each page, serving as filler items for the alternate experiment.

Experiment 1 aimed to investigate whether the position of the subject in the given sentence influenced the position of the corresponding emoji in the resulting sequence. For this purpose, five sentences were constructed with the sentence structure Subject-Predicate-Adverbial (S-P-A) and another five with the structure Subject-Predicate-Object (S-P-O). Additionally, manipulated versions of these sentences were created by inverting the structure to Adverbial-Predicate-Subject (A-P-S) or Object-Predicate-Subject (O-P-S). For instance, group A received the sentence *Die Sonne scheint warm* ('The sun shines warmly'), while group B received its inverted version *Warm scheint die Sonne* ('Warmly shines the sun'). Note that both the default (subject-first) word order and the reverse are equally grammatical in German. The sentence variants were evenly distributed between the two groups so that both group A and group B received five sentences in their original word order and five in the reverse order, minimizing the impact of potential non-measurable factors within a single group.

Experiment 2 examined the semiotic interpretation of human-representing emojis within the subcategory "Activities and Sports" (part of the "People" category²). The central research question was whether these emojis are interpreted solely as depicting human actors or whether they can also denote the activity independently of the human agent. Ten simple sentences were constructed in which humans performed activities that could be represented by emojis from the specified subcategory. Each of these sentences had an alternate version in which the human agent was replaced by an animal. As in Experiment 1, both versions were evenly distributed across the two groups, with each group receiving five sentences in their original form and five in their manipulated form. For instance, group A saw the sentence *Die Frau spielt Golf* ('The woman plays golf'), while group B received the sentence *Der Pinguin spielt Golf* ('The penguin plays golf').

It is worth noting that, due to technical constraints, nouns referring to humans were exclusively presented in either female or male forms. The reason is that many human-representing emojis were only recently

2. For further reference, see <https://emojipedia.org/peopleactivities-sports>. Additionally, there is a separate category called "Activity" including representations of sports gear without depicting humans, see <https://emojipedia.org/activity>.

updated to include gender-neutral versions in the Unicode standard, and these are not properly displayed in some browsers³.

3.1.3. *Data*

The raw data for the analysis is provided by SoSci Survey, which assigns a unique case number to each participant and records all responses as well as the time required to complete the tasks. This file was further processed by filtering incomplete datasets, resulting in 27 datasets from group A and 22 datasets from group B being included in the analysis. Since the demographic composition of the two groups is very similar (see the full data table for details), we consider the difference in group sizes to be negligible.

Key relative frequencies relevant to the experiment were also recorded in the table. It should be noted that SoSci Survey automatically converts responses into text format. To avoid errors, we conducted a pre-test in which we entered all emojis to verify their respective conversion and ensure correct assignment.

For illustrative purposes, selected examples will be presented in 3.2 and 3.3 as emojis rendered using the current operating system iOS 17.4 by Apple. This approach aims to accurately reflect how these emojis are displayed in practice.

The experiment data and analyses are available at the address <https://osf.io/qhgx3/>.

3.2. Results and Discussion 1: Altered Subject Positions

To analyze the response data, we categorized the obtained emoji sequences for each sentence into three result categories:

1. **Consistent Subject Position:** This category includes all emoji sequences that consist of at least one representation of the subject and one additional sentence constituent, with the subject-representing emoji(s) placed in the same position as the subject in the given sentence.
2. **Shifted Subject Position:** This category encompasses all emoji sequences that consist of at least one representation of the subject and one additional sentence constituent, but where the subject-representing emoji(s) were placed in a different position than the subject in the given sentence. This includes cases where the sequence reflects an inverted structure compared to the given sentence, as well as sequences where multiple sentence constituent positions do not correspond to the order in the original sentence.

3. See <https://unicode.org/emoji/charts/full-emoji-list.html>

3. **Alternative Representation:** This category covers all emoji sequences that either represent only a single sentence constituent or do not include any emoji that can be interpreted as representing the subject of the given sentence. This does not necessarily imply that the task instructions were ignored; in some cases, patterns can still be observed for certain sentences.

It is important to note that, across all result categories, not all sequences align with the categorical grammar framework of Cohn, Engelen, and Schilperoord (2019). In many cases, assigning sequences to a specific grammatical structure or distinguishing between patterns is challenging. In the following sections, some of these representations will be further analyzed and interpreted to gain deeper insights into the observed phenomena.

3.2.1. *Subject-Predicate-Adverbial*

After categorizing the responses into the three result categories, the frequency distribution for sentences with the structure S-P-ADV is shown in Figure 1, with the numbering of the sentences corresponding to that in the questionnaire appendix. The addition of “i” denotes the inverted form of the sentence with the structure ADV-P-S.

In this section, the results for these sentences will be presented, highlighting the frequency of emoji sequences that correspond to each of the three categories (consistent subject position, shifted subject position, and alternative representation). This analysis will provide insights into how the participants adapted the sentence structure when translating it into emojis, and whether the inversion of the sentence order affected the placement of the subject-representing emojis.

It is immediately apparent that the subject position is mostly retained with values between 77.8% and 90.9% when it appears at the beginning of the given sentence. In contrast, when the sentence structure is inverted, the subject position is more frequently shifted, with values ranging from 40.7% to 72.7%, although with greater variability in the data. This significant difference between the data from the respective sentences and their inverted variants indicates that participants did not simply adhere to the syntax of the given sentence. Furthermore, it is unlikely that the order within the emoji sequence was chosen entirely arbitrarily; for example, shifting the subject position when it appears at the beginning of the sentence occurs at a frequency of at most 7.4% in Sentence 5, making this a rather exceptional case.

The frequency of alternative representations varies across sentences, ranging from 4.5% to 29.6%. In line with the assertion that performance does not directly reflect competence (Chomsky, 1985, p. 4), it cannot be concluded from the present data that emojis inherently possess the

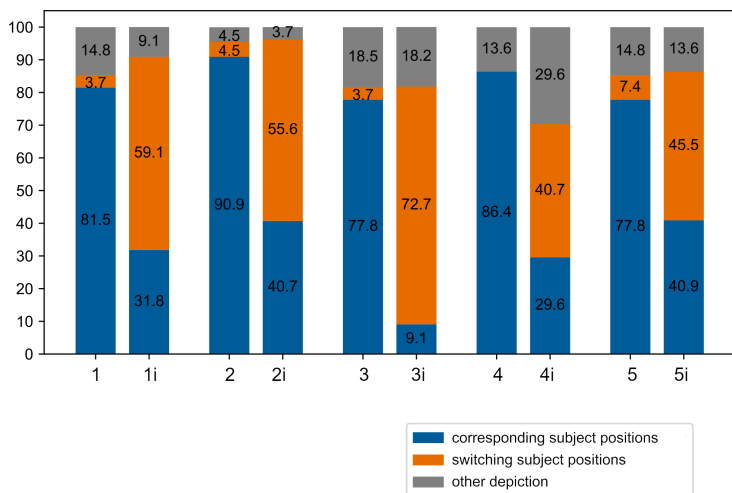


FIG. 1. Results of Experiment 1a: Subject position in emoji sequences for S-P-ADV (unmarked) and ADV-P-S (inverted, indicated by “i”) syntax

grammatical property of positioning the subject at the beginning of the sentence. It is also plausible that participants oriented themselves towards the more familiar syntax; after all, starting a main clause with the subject is typical for German.

In the following, we will discuss several examples that illustrate the responses obtained.

- (1) *Die Sonne scheint warm.*
 ‘The sun shines warmly.’

- | | | |
|----|--------|------------------------|
| a. | ☀️🔥 | (188/A1*) ⁴ |
| b. | ☀️😬 | (106/A1*) |
| c. | 👉☀️🔥🌡️ | (140/A1) |
| d. | ☀️ | (178/A1*) |

Emoji sequences like those in (1a) and (1b) were frequently used, while (1c) represents a single exception. A word-for-word “translation” was therefore more likely to be rejected by participants, although it demonstrates that such translations are not impossible. This contradicts the

4. The numbers identify the response in the raw data table described earlier. The first part of the identifier indicates the number for the row in the document. The second part provides the SoSci Survey question identifier, which can be found in the headers of the respective columns. Emoji sequences that appear multiple times in the responses (disregarding any skin tone modifications) are marked with an asterisk.

assumption that a definite article cannot be represented by emojis (Dürscheid and Siever, 2017).

In the inverted version of the sentence, examples (1a) and (1b), or the same sequences in reverse order, were also observed. It is noteworthy that, in addition to representing the subject, emojis were used to depict the adverbial adjective *warm*, with the predicate being entirely omitted. The meaning had to be inferred from the interaction of the two emojis. This tendency can be observed for the other sentences as well.

Alternative representations, such as in (1d), include only a single sun emoji. It is conceivable that such a representation was chosen because the sun inherently has the properties of shining and radiating warmth. Therefore, adhering to grammatical structures may be seen as unnecessary under certain conditions. This aligns with Pappert's (2017) assertion that emojis serve an *economic function*, simplifying communication by reducing the need for complex syntactic constructions.

(2) *Laut singt der Vogel.*

'Loudly sings the bird.'

a.  

(120/A2*)

b.  

(154/A2)

Larger differences in representation are observed in the results for sentence 2i. The frequent combination of a speaker and musical notes, as seen in (2a), suggests that, unlike in sentence 1, a combination of two emojis (or even one) is insufficient to convey the meaning of the sentence. This could be due to the fact that a sequence like (2b) may capture the semantic features [+bird] and [+loud], but the melodic property of singing is completely omitted—unlike with the sun, where even a single emoji can evoke an associated property like warmth.

This highlights that some actions or properties require a more nuanced representation through emojis, as certain meanings (such as the melody in singing) cannot be captured by basic visual signs. This suggests that the representation of more complex or dynamic concepts may necessitate the use of multiple emojis to fully convey the intended message, as opposed to simpler concepts, which can often be represented with a single emoji.

(3) *Der Vogel singt laut.*

'The bird sings loudly.'

(82/B2)

The emoji sequence depicted in (3) is the only one where the bird emoji is not congruently positioned at the beginning, following the subject position as in the original sentence. In addition to the repetition of musical notes as a symbol for volume, the microphone can be interpreted as representing the predicate. We argue that this is not merely a

positional shift of the subject based on grammatical properties. Rather, the emojis have been arranged so that the bird directs its gaze toward the microphone, and the musical notes visually originate from the microphone and fly away from the bird.

This raises a challenge in relation to the hierarchy established by Cohn, Engelen, and Schilperoord (2019): Cohn, Engelen, and Schilperoord (*ibid.*) emphasize that a *one-unit grammar* does not mean a single emoji is used, but rather it can also apply when multiple emojis are used to create a larger *whole image*. In principle, this applies here as well; however, the repetition of the musical note emoji introduces a higher grammatical category within the linear grammar hierarchy. One might assume that by embedding one category within another, the sequence as a whole would be classified as an *embedded sequence*, representing the highest grammatical category. However, a significant difference from the examples of Cohn, Engelen, and Schilperoord (*ibid.*) lies in the fact that individual sequences are not combined here. While the three musical note emojis represent a *repetition*, they are not combined with a second grammatical form, which consists of the microphone and bird emojis. Instead, the whole image emerges only through the combination of all five emojis.

This suggests that a category above in the hierarchy does not necessarily dominate or subsume a lower category. Additionally, this sequence does not seem to follow higher grammatical structures, and it is likely that no intention to form sentence constituents was present. The idea that such a case, which lacks the recognition of basic grammatical structures, would fall into the same embedded sequencing category as the examples from Cohn, Engelen, and Schilperoord (*ibid.*), where fundamental grammatical structures are identifiable, seems inadequate to us.

This analysis challenges the application of Cohn et al.'s hierarchical model to more complex and abstract emoji sequences, as the ordering and usage of emojis here suggest a different type of syntactic or visual organization that does not strictly adhere to the basic structures identified in their framework.

(4) *Schnell fährt das Auto.*

'Fast drives the car.'

a. 🚗💨

(104/B3*)

b. 🚗💨💨💨

(78/B3*)

Sentence 3i shows the lowest frequency of 9.1% for maintaining the subject position. Emoji sequences such as those shown in (4), represent the largest share of the depiction. The same holds true for sentence 3. Here, as in example (3), it is evident that the sequence is less oriented toward sentence constituents and more toward a pictorial representation, following a *one-unit grammar* (*ibid.*), in order to form an *affixation*. Once again, the direction of the car and dust cloud emojis plays a significant role. It

can be assumed that if these were presented mirror-reversed in the emoji keyboard, the preferred representation would likely involve an inversion of the sequences in (4).

This would also highlight a feature that continues to distinguish emojis from writing systems of both alphabetic and logographic languages. We suspect that a mirror image of German alphabetic characters would disrupt reading flow at least to some degree and doubt that such a reversal would result in a preference for the syntactic swapping of individual words or morphemes within a sentence.

In contrast to written language, where the inversion of individual characters or words can lead to confusion or difficulty in reading, the flexibility and potential for visual play in emoji sequences do not seem to face the same constraints. Emojis function in a more visual or pictorial domain where spatial and directional manipulations of symbols may not necessarily disrupt comprehension.

- (5) *Schnell fährt das Auto.*
'Fast drives the car.'



(74/B3*)

- (6) *Hoch fliegt das Flugzeug.*
'High flies the airplane.'



(264/A4*)

Examples (5) and (6), consistent with (1d), show that a single emoji can also be perceived as sufficient when it semantically encompasses all the properties conveyed in the sentence. For example, a race car is attributed the inherent property [+can drive fast], while an airplane is attributed the property [+flies high].

We would also like to argue that the proportion of other representations for sentences 4 and 4i is relatively high due to the use of a single airplane emoji, as the design of the emoji visually emphasizes the direction of flight. This visual cue reinforces the semantic message of the sentence, which may make a single emoji appear adequate to represent the entire meaning. By highlighting specific characteristics such as speed or altitude, the emoji can provide a complete, contextually appropriate "translation" of the sentence, thereby making it a functional representation in the context of emoji-based communication.

3.2.2. Subject-Predicate-Object

Following the division into the three result categories, the frequency distribution for sentences with the structure S-P-O is shown in Figure 2, where the sentence numbering corresponds to the one in the questionnaire in the appendix. The suffix "i" represents the inverted form of the sentence with the structure O-P-S.

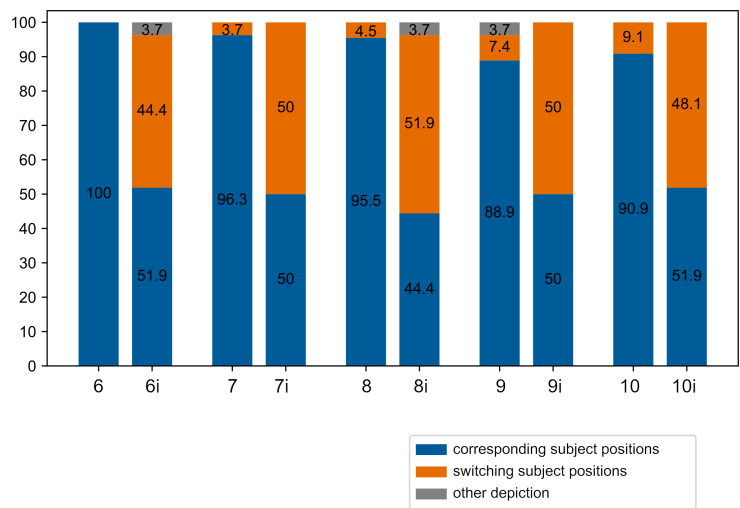


FIG. 2. Results of Experiment 1b: Subject position in emoji sequences for S-P-O (unmarked) and O-P-S (i = inverted) syntax

In comparison with the results from 3.2.1, it is evident that the proportion of other representations, often involving just a single emoji, is significantly lower, with values ranging from 0% to 3.7%. This supports my assumption that these were not chosen arbitrarily or due to ignorance, but rather because the adverbial adjective describes an inherent property of the subject and does not necessarily require its own representation. However, for sentences with an object, it seems necessary to provide an explicit representation of the object. It is observed that the alignment of the subject position, on average, exceeds 90% when the subject is at the beginning of the sentence, indicating that there is neither significant rejection nor arbitrary changes in this case. For sentences where the subject is at the end of the sentence, there is nearly a 50%-50% ratio on average when it comes to maintaining the subject position, with only minor variability. This initially suggests that there is no consensus on a fixed rule. However, an analysis of selected examples in 3.2.1 showed that the syntax does not necessarily depend on the order of the sentence components. Rather, the direction of representation of certain emojis, such as in examples (3) or (4), seems to have a significant influence on the participants' choices.

- (7) *Die Hummel liebt Blumen.*
'The bumblebee loves flowers.'
- a. 🐝❤️🌹🌺🌸 (228/A9)
- b. 🐝😍🌸🌺🌹🌻 (162/A9)

The fact that the direction of emojis is not a sufficient factor in determining sentence syntax is demonstrated by examples (7). In these cases, the bee emoji explicitly looks away from the flower emojis, yet 88.9% of participants chose a syntactically comparable representation. It is also notable that for the word *flowers*, in most cases, a *semantic list* (Cohn, Engelen, and Schilperoord, 2019) was created, which was embedded into the syntactic structure of the sentence. This highlights both the plural and the playful component of the selection of flower emojis, which Papert (2017) refers to as a *ludic function*. A bee emoji facing the flower emojis could be congruent with the syntax in sentence 9i. However, 50% of participants chose against this and created a sequence comparable to those in (7). This value is similar to the other sentences 6–10, even though no emoji was chosen that is oriented to the left or right.

We would like to propose the hypothesis that the direction of an emoji is particularly important when an affixation like in (4) is formed, but less so when a whole image is created. This could be tested in future experiments.

- (8) *Der Weihnachtsmann liest ein Buch.*
 ‘Santa Claus reads a book.’

- a. 🧑🏻📖 (74/B8*)
 b. 📖👁️🧑🏻 (254/B8)

Here, the emoji sequence in (8a) exemplifies the most frequently chosen representation of the sentence (68.2%, regardless of skin tone modification). Compared to (8b), the only instance in which the subject’s position was swapped, the representation of the predicate is omitted. This may be related to the fact that an open-book emoji intuitively associates with reading, and also that the eye emoji directs its gaze to the left. To avoid ambiguity, the subject position would need to be swapped, which is less preferred. In sentence 8i, a representation with the eye emoji while maintaining the subject position appears only 14.8% of the time, suggesting that economy, driven by semantic properties, likely has a stronger influence on the responses.

- (9) *Das Baby malt einen Regenbogen.*
 ‘The baby paints a rainbow.’

- a. 🧑🏻🎨🌈 (232/B10*)
 b. 🧑🏻🖍️🌈 (230/B10*)
 c. 🧑🏻🎨🌈 (216/B10*)

The various examples for sentence 10 show an orientation toward the syntactic structure of S-P-O. In only 9.1% of the cases, a representation of the predicate was omitted. It is assumed that a rainbow does not have the prototypical property of being painted, which makes a representation

smaller than a *three-unit sequence* (Cohn, Engelen, and Schilperoord, 2019) less preferable for reasons of economy.

In conclusion, it can be stated that, in the context of the experiment, emoji sequences appeared at all levels of the hierarchy established by Cohn, Engelen, and Schilperoord (*ibid.*). A preference based on sentence structure could not be demonstrated. Rather, specific properties, such as gaze direction and prototypical associations, seem to be decisive for the “translation” of the sentence. Considering the guidelines given for future emojis, the establishment of emojis as an independent language seems highly unlikely and currently unattainable. However, their use as lexical units is possible, as argued by Dürscheid and Siever (2017) and Weissman, Engelen, Baas, and Cohn (2023), among others. In response to the initial research question, we can conclude that a subject position at the beginning of the emoji sequence is preferred when it aligns with the subject position in the given sentence. When the subject is at the end of the sentence, no clear preference is observed for the subject’s position within the response. Maintaining and shifting the position of the subject-representing emoji are equally preferred or rejected on average. Some emoji sequences exist at a lower grammatical level, which does not follow the formation of sentence constituents. The semiotic qualities of the emojis also influence the order within a sequence. We therefore assume that emojis provide incentives to forgo strictly fixed grammatical structures.

3.3. Results and Discussion of Exp. 2: Semiotics of Activity Emojis

As already shown in the examples from 3.2, the semiotic quality of emojis is not easily determined. Examples (1d), (5), and (6) demonstrate that an emoji can represent not only individual concepts of entities but also entire propositions, although this may not necessarily be the preferred application. For activity emojis depicting people (hereafter abbreviated as HAEs), it is conceivable that these can represent an entire proposition or only the activity itself. It is also important to examine whether this changes in different contexts, thus suggesting that a semiotic association (in the sense of Saussure, 2011) juncture to represent an activity (Grosz, Kaiser, and Pierini, 2021); however, no distinction is made between HAEs and emojis from the separate category of “activities” that do not depict people. The results of this study can build upon these findings.

For the analysis of the response data, we have divided them into the following result categories:

1. Subject Representation + Human Activity Emoji: This category includes all emoji sequences that, in addition to an HAE, contain at least one other emoji representing the agent of the sentence.

2. **Subject Representation + Other Representation of Activity:** This includes all emoji sequences that feature at least one emoji representing the agent (hereafter abbreviated as AR) and at least one other emoji depicting the activity, without using an HAE.
3. **Only Human Activity Emoji:** This category, which only appears in 3.3.1, includes emoji sequences consisting solely of a single HAE or its repetition.
4. **Other Representation:** Emoji sequences that do not include an HAE and either lack a representation of the agent or the activity are categorized here.

Occasionally, there were responses that did not adhere to the task and included additional characters alongside emojis. These are categorized as “not adhering to the task” and are included in the graphs in 3.3.1 and 3.3.2.

3.3.1. *Humans as Agents*

Based on the classification into the previously mentioned result categories, the frequency distribution for sentence variants with a human agent is shown in Figure 3, with the sentence numbering corresponding to the questionnaire in the appendix.



FIG. 3. Results of Experiment 2a: Usage of human representing activity emojis in sentences with human agents

Significant deviations in the frequencies of the first three result categories are observed. The frequency of an emoji sequence with AR (Agent Representation) combined with at least one HAE (Human Activity Emoji) averages 39.2%, with a considerable spread ranging from 14.8% to 70.4%. A comparable average is observed for the representation using only HAE, at 38.2%, although this category exhibits an even larger spread, with values ranging from 13.6% to 72.7%. A lower average of 19.8% is found for emoji sequences with AR combined with a representation of the activity without HAE, with values ranging from 0 to 54.5%. Other result categories appear only rarely, with a maximum value of 9.1% for sentence 10. Notably, despite the significant spread in the values, certain pairs exhibit very similar frequency distributions, such as sentences 11 and 14, sentences 13 and 15, and sentences 16 and 18. We will now discuss in the following examples whether patterns can be identified that point to underlying reasons for these results.

- (10) *Peter schwimmt.*
'Peter swims.'

- a.  (74/B16*)
b.   (160/B16)

The most common representation of sentence 16 is shown in (10a). The chosen HAE depicts a male person, interpreted as *Peter*, wearing a swim cap and goggles, engaged in a swimming motion in the water. Less frequent representations, such as in (10b), place an emoji before the HAE, which could be interpreted as indicating gender or suggesting that the person depicted is not arbitrary. An identical distribution of responses based on the selected result categories is also found in the following sentence:

- (11) *Laura tanzt.*
'Laura dances.'

- a.  (258/B18*)
b.   (70/B18)

The HAE chosen in (11a) depicts a female person, interpreted as *Laura*, wearing a flowing dress and adopting a dancing posture. Less frequent emoji sequences with a preceding AR also appear, which might indicate gender or a specified name. Sentences 16 and 18 share a distinctive feature compared to all other sentences: both consist solely of a subject and predicate, each represented by a single word. The highest response rate with exclusive HAE representation, 72.7%, in sentences 16 and 18 correlates with this structure.

- (12) *Die Frau jongliert.*
'The woman juggles.'

- a.  (78/B12*)
 b.  (216/B12*)

While the representation of sentence 12 through a combination of AR and HAE, as seen in (12a), is preferred with a frequency of 55%, the representation through a single HAE, at 45%, is similarly common. Since the juggler emoji marks the feminine attribute primarily through the hair partially covered by the hat—unlike the dancer emoji in (11)—this might explain the slightly higher rate of the additional AR. It is also worth noting that there are no responses to sentences 12, 16, and 18 without an HAE.

- (13) *Die Frau fährt Ski.*
 ‘The woman skis.’

- a.  (104/B14*)
 b.  (236/B14*)
 c.   (222/B14*)

Sentence 14 contains the same subject as sentence 12 and shows similar results, but with a significantly differing frequency distribution. The representation with a single HAE, as seen in (13b), is only present in 18.2% of cases, while emoji sequences consisting of both AR and HAE are present in 68.2%. Another representation of the activity using the ski equipment emoji is also relatively rare, at 13.6%. This suggests that an HAE is used to represent the activity itself, and additional AR is only necessary when certain properties, such as gender, are not identifiable in the HAE. The ski HAE in (13a) and (13b), unlike the previously discussed HAEs, is not available in multiple gender-marked variants in Unicode, which may explain the high rate of emoji sequences as in (13a).

At this point, we would like to reject the previously made assumption that an AR placed before the HAE functions to distinguish between specific and unspecified entities. While *Peter* and *Laura* in sentences 16 and 18 are more specific than *the woman* in sentences 12 and 14, a preceding AR occurs significantly more often in sentences 12 and 14, suggesting it can be interpreted more as a gender marker.

- (14) *Leon und Florian halten Händchen.*
 ‘Leon and Florian hold hands.’

- a.  (84/A17*)
 b.    (180/A17*)
 c.    (246/A17)

Sentence 17 exhibits a unique frequency distribution of result categories. While 37% of the emoji sequences are represented by a single HAE as in (14a), only 14.8% consist of a combination of HAE and AR. The largest proportion, 44.4%, consists of a combination of emojis representing the

agent and a representation of the activity without HAE, as seen in (14b) and (14c). It is difficult to clearly identify the reasons for this. Representations similar to (14b) for "Leon" and "Florian" always feature the same emoji, which excludes a more precise differentiation of the individuals through distinct heads. Perhaps the depiction of multiple people within one emoji also leads to this being less recognized and therefore less frequently chosen. Example (14c) forms an interesting exception, where the names are presented playfully: the lion emoji for "Leo" and the tulip emoji for "Florian", likely referencing the word "Flora." This is an example of wordplay, as addressed by Gawne and McCulloch (2019), the meaning of which can be difficult, if not impossible, to disambiguate without context.

In general, it is noticeable that the depiction of specific names posed problems. The appearance of sequences like (14c) shows that participants did not consider this aspect irrelevant. For sentences 17, 19, and 20, there are even occasional responses in which the task was ignored, and names were written out without emojis. This illustrates that not only can complex concepts not be adequately expressed with emojis (Dürscheid and Siever, 2017), but even very simple yet specific ideas cannot always be properly conveyed. Therefore, the experiment, as already mentioned, adapted the sentences mostly to the available visual vocabulary (Cohn, Engelen, and Schilperoord, 2019) of emojis.

3.3.2. *Animals as Agents*

In order to assess the semiotic qualities of HAEs, it is essential to first examine and discuss the results of their use in a context without human agents. To confirm the hypothesis that HAEs can denote activity independently of a human context, it would be expected that these appear with similar frequency in the manipulated sentence variants 11t to 20t. For sentence variants with animal agents, the frequency distribution shown in Figure 4 applies, with the sentence numbering matching that in the questionnaire in the appendix. The suffix "t" marks that these are the manipulated versions of the sentences without the addition, in which the agent has been replaced by an animal.

Even when humans are replaced by animals as agents, no consistent pattern emerges regarding the form of representation. Combinations of AR and HAE are used with great variation, ranging from 14.8% to 88.9%, while combinations of AR and representation of the activity without HAE occur with frequencies ranging from 11.1% to 95.5%. A representation consisting solely of HAE does not occur at all. This observation is consistent with the earlier conclusion that an emoji is considered necessary as an additional AR if it expresses important properties that are not made visible by the HAE.

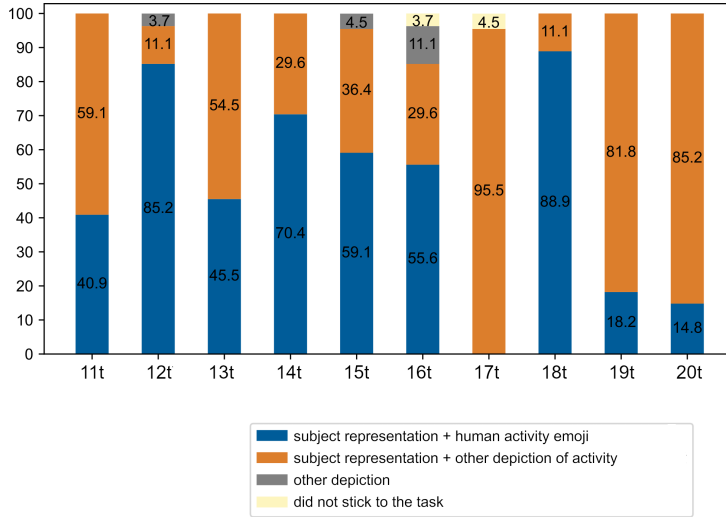


FIG. 4. Results of Experiment 2b: Usage of human representing activity emojis in sentences with animal agents

Unlike the variants with human context, similar correlations of frequency distributions between the sentences 11t and 14t, 13t and 15t, and 16t and 18t cannot be observed. According to my assumption that the correlation between the values of sentences 16 and 18 is due to these sentences being the only ones consisting of just two words, it explains why this correlation is absent in sentences 16t and 18t. However, with animals as agents, similar frequency distributions can be found in sentence pairs 11t and 13t, 12t and 18t, as well as 19t and 20t. A more detailed examination of individual representations follows.

(15) *Der Affe jongliert.*
 ‘The monkey juggles.’

- a.  
 b.  

(204/A12*)

(124/A12)

Example (15a) shows the most frequently given response. Here, the monkey emoji is used as an AR combined with the HAE of the juggling person, which was also used in sentence 12. A plausible assumption is that the HAE is lexically located as a sign for *juggling*, independent of the human characteristic. In contrast, emoji sequences consisting of AR and representation of the activity without HAE, such as in (15b), were used only 11.1% of the time, and present a greater challenge for disambiguation: without context, it is unlikely that this sequence could be traced back to the provided sentence.

(16) *Der Bär tanzt.*
 ‘The bear dances.’

- | | | |
|----|---|-----------|
| a. |   | (144/A18) |
| b. |   | (218/A18) |
| c. |    | (184/A18) |

Sentence 18t, with a frequency distribution similar to that of sentence 12t, shows the same pattern with the frequent responses seen in (16a) and (16b): the bear as AR and the dancing people as HAE. The fact that both male and female HAEs occur (and, unlike the juggling person, there is no gender-neutral emoji for this) supports the assumption that the HAE is lexically located for performing the activity. No correlation can be observed between the perceived gender of the dance emoji and the gender provided by the participant who used the emoji in their response. This also reflects the results of Grosz, Kaiser, and Pierini (2021), who view activity emojis as free adjuncts, unlike face emojis, which are not primarily used for referencing the self. Among the less frequent responses (11.1%) is (16c). Here, similar to (15b), a context-sensitive visual puzzle is created, where it is assumed that disambiguation is made more difficult.

(17) *Das Känguru spielt Basketball.*
 ‘The kangaroo plays basketball.’

- | | | |
|----|---|------------|
| a. |   | (160/B19*) |
| b. |   | (71/B19*) |

A significantly different frequency distribution can be observed in sentence 19t. Emoji sequences consisting of AR and HAE, as shown in (17a), make up only 18.2% of the responses. The variant, preferred with 81.2%, consisting of AR and representation of the activity without HAE, is shown in (17b), the most frequent response. Since this distribution is not an exception, the question arises as to why the use of the HAE is more often avoided. One possible reason we see is that the representation in (17b) does not, like in (15b) or (16c), lead to difficult disambiguation. We suspect that the prototypical association with a basketball is playing, whereas the ballet shoes depicted in (16c), as clothing, first make one think of putting them on or wearing them. And although pins are frequently used with balls for juggling, the ball in the bowling emoji in (15b) more strongly evokes the association with bowling. For sentence 19t, it could even be argued that the use of the HAE in (17a) leads to more difficulty in disambiguating, as a joint game involving a kangaroo and a human could be interpreted. The high rate of HAE in individual sentences in the animal context may therefore not be due to a semiotic connection between the HAE and the activity, but rather a lack of alternative available emojis. However, this also suggests an attempt to use

emojis as lexical units and to avoid constructing picture puzzles that are more difficult to disambiguate.

(18) *Die Katze fährt Boot.*

‘The cat goes boating.’

- | | | |
|----|---|------------|
| a. |   | (178/A20*) |
| b. |   | (204/A20) |
| c. |    | (144/A20) |

The responses to sentence 20t paint a picture similar to that of sentence 19t: The representation with 14.8% preference, consisting of AR and HAE, as shown in (18a), suggests that disambiguation is made more difficult. A cat riding a boat is not a typical event; thus, the HAE in (18a) can also be interpreted as a human riding away from the cat in the boat. This assumption is reinforced by the interpretation of other responses: Unlike the responses for most of the other sentences, many participants seemed to feel that a representation with just two emojis, as shown in (18b), was not sufficient. Various emoji sequences, exemplified in (18c), contain a third emoji to connect the cat and boat emojis, as this is presumably necessary to express the relationship between “cat” and “boat.”

(19) *Der Hase fährt Ski.*

‘The rabbit skis.’

- | | | |
|----|---|------------|
| a. |   | (158/A14*) |
| b. |   | (106/A14*) |

A frequency distribution with values of 70.4% for the combination of AR and HAE and 29.6% for the combination of AR and a representation of the activity without HAE is observed for sentence 14t. It can be argued that the emoji sequences shown in (19) have a similar degree of difficulty in their disambiguation. While the HAE in relation to the rabbit emoji can be interpreted as an activity performed by the rabbit, an interpretation is also possible where a human skier is approaching the rabbit. Similarly, a prototypical association with the ski equipment could be that the rabbit is merely wearing it or is depicted as moving away from it due to the direction of its gaze. In this case, a preferred representation could be due to individual aesthetic reasons. The most frequent answers for sentences 15t and 16t show a similar pattern. Here too, a representation with HAE is more commonly seen, but a significant portion, 36.4% and 29.6%, respectively, of answers use AR and non-HAE to represent the activity. Answers from both result categories seem equally plausible (see the full data as referenced in 3.1.3), and the exact reasons for the preference can also be attributed to immeasurable factors.

In conclusion, it can be stated that HAEs are frequently used both in animal and human contexts. When characteristics such as gender

marked by a word are visually represented by an HAE, it is seen as unnecessary to form an emoji sequence with an additional AR. In the case of animals as the agent of the sentence, the frequency of HAE usage varies depending on the alternative emojis available. There are cases where an HAE either complicates, simplifies, or likely reflects individual preference for non-measurable reasons. Arguments for considering emojis as an independent language based on *langue* and *parole* (Arafah and Hasyim, 2019; Duque, 2018) therefore seem inadequate and linguistically unsustainable. Regarding the initial question about HAEs, we conclude the following: Clearly defined semiotic relations for HAEs in human contexts can be assumed initially, but when the human characteristic of the agent changes, they can no longer be clearly proven. The individual visual properties of the HAE, as well as prototypical associations between agent and activity, influence the representation of emoji sequences.

3.4. Limitations

Limitations arise from the nature of the study. It must be considered that the results of such an experimental, anonymous online study cannot provide sufficient evidence for generalizable rules regarding the use of emojis. A larger sample size or deliberately selected groups, sorting participants by age, gender, or other demographic data, may yield different results. Due to the anonymous online format, it cannot be ruled out that an individual may have participated multiple times. The interpretation of the results is not always clear-cut, and the reasons for the use of certain emojis and their order in the sequence may vary individually due to immeasurable factors. In contrast to ASCII symbols like emoticons, emojis are not arbitrarily composed products of multiple characters (cf. Pavalanathan and Eisenstein, 2016), but rather individual characters that are firmly embedded in the operating system. Since 2010, they have been encoded in the international character standard Unicode. The representation varies depending on the chosen software, analogous to the use of different fonts, in which the fundamental characteristics of an emoji are maintained so that it can be as easily recognizable as possible. However, it cannot be ruled out that the platform-specific display of individual emojis might be a small but not entirely negligible factor in their disambiguation and, thus, in their usage. It is also assumed that not all participants are familiar with all emojis. Since Unicode provides a four-digit number of emojis on the keyboard, unexpected or shorter sequences may arise for motivational or time-related reasons. Finally, it should be noted that a different selection of sentences could potentially produce different results. Further experiments of this kind are necessary for verification.

4. Conclusion

The aim of this work was to investigate the current linguistic status of emojis. Specifically, we investigated their grammatical and semiotic qualities. Cohn, Engelen, and Schilperoord (2019) established a hierarchy of grammatical complexity for emoji sequences, which shows that basic grammatical constructions are possible. Recent research by Weissman, Engelen, Baas, and Cohn (2023) further indicates that emojis are currently undergoing a lexicalization process, correlating with their category, making context-independent disambiguation more difficult to almost impossible. The empirical study we conducted followed up on this research and asked participants to convert given simple sentences into emoji sequences using free text input in randomized order. Two research questions with ten sentences each were examined. A random assignment into two groups with manipulated sentence variants produced measurable results.

Experiment 1 investigated how the subject position in the given verbal sentence affects the position of the subject-representing emoji in the emoji sequence. The results showed that the subject position is largely preserved when the subject is at the beginning of the sentence (canonical subject-predicate order). When the subject is placed at the end of the sentence (inverse predicate subject order), participants deviate from the structure as often as they adhere to it. The variability between the results of individual sentences is small. The hypothesis that emojis already follow established grammatical structures is therefore unlikely. Additionally, many emoji responses did not adhere to the hierarchical construction of phrases but formed sequences that can be classified at a lower grammatical level. Visual factors such as the direction of an emoji's gaze influenced its position within the sequence. Prototypical associations with the entity represented by the emojis also influenced the decision about whether an (additional) emoji was necessary to represent a word from the original sentence. A notably frequent occurrence was the lack of a separate representation of the predicate.

Experiment 2 examined the semiotic qualities of human-representing activity emojis (HAE). The results primarily found a preferred selection of HAE in the human context. When the agent is assigned an important characteristic, such as gender, which is not visually represented by the HAE, participants added additional emojis to represent the agent. When the agent is replaced by an animal, the frequency distribution of emoji sequence types deviates from those of the sentence variants with a human agent. The use of HAE is rejected when an alternative representation of the activity (e.g., a salient object) leads to less difficulty in disambiguating the sequence. Again, individual visual properties of emojis, as well as prototypical associations with the entities they represent, have an influence.

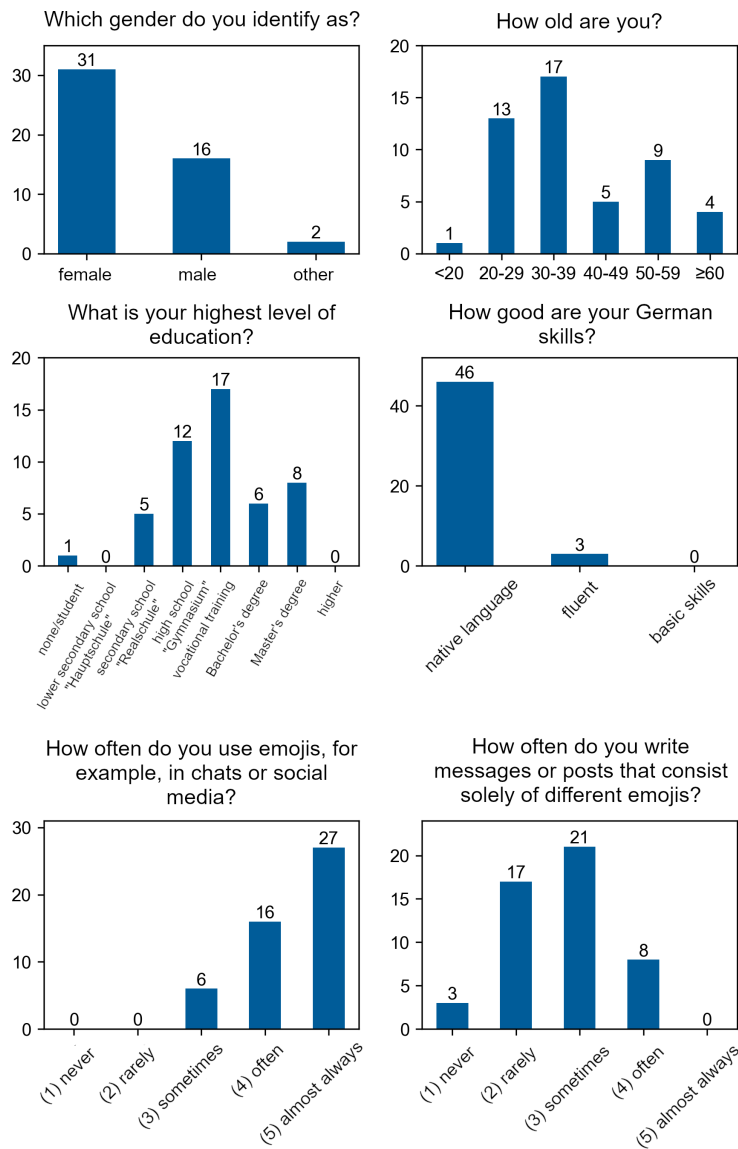
Our results show that while it is possible to represent simple sentence meanings purely with emojis, these representations show little structure and systematicity, and are often either dependent on the underlying (textual) sentence before translation, or mere depictions of the described situation. Importantly, visual characteristics of the emoji characters, such as their spatial orientation and whether certain properties are visually salient or not, are sometimes more significant for emojis' use in our participants' translations than any abstract 'meaning' assigned to them via their official names. It thus seems that the majority of our participants select (object and activity) emojis not as a 'writing system' for linguistic output (even when asked to do so in an experiment), but rather via distinct processes.

References

- Ai, Wei et al. (2017). "Untangling Emoji Popularity Through Semantic Embeddings". In: *Proceedings of the International AAAI Conference on Web and Social Media* 11.1, pp. 2–11.
- Androutsopoulos, Jannis (2003). "Online-Gemeinschaften und Sprachvariation. Soziolinguistische Perspektiven auf Sprache im Internet". In: *Zeitschrift für germanistische Linguistik* 31, pp. 173–197.
- (2020). "Digitalisierung und soziolinguistischer Wandel". In: *Deutsch in Sozialen Medien. Interaktiv—multimodal—vielfältig*. Ed. by Konstanze Marx, Henning Lobin, and Axel Schmidt. Berlin, Boston: De Gruyter, pp. 75–94.
- Arafah, Burhanuddin and Muhammad Hasyim (2019). "The Language of Emoji in Social Media". In: *KnE Social Sciences, The 2nd Annual International Conference on Language and Literature (AICLL 2019)*, pp. 494–504.
- Chomsky, Noam (1985). *Aspects of the Theory of Syntax*. [original work published 1965]. Cambridge, Massachusetts: The M.I.T. Press.
- Cohn, Neil, Jan Engelen, and Joost Schilperoord (2019). "The grammar of emoji? Constraints on communicative pictorial sequencing". In: *Cognitive Research: Principles and Implications* 4, Article 33: 1–18.
- Cohn, Neil et al. (2018). "Are emoji a poor substitute for words? Sentence processing with emoji substitutions". In: *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, pp. 1524–1529.
- Duque, Diana (2018). "I heart [emoji]: Understanding the effects of a new language of self-expression". In: *PLOT(S) Journal of Design Studies* 5, pp. 54–66.
- Dürscheid, Christa and Dimitrios Meletis (2019). "Emojis: A Grapholinguistic Approach". In: *Proceedings of Graphemics in the 21st Century, Brest 2018*. Ed. by Yannis Haralambous. Brest: Fluxus Editions, pp. 167–183. DOI: <https://doi.org/10.36824/2018-graf-duer>.

- Dürscheid, Christa and Christina M. Siever (2017). "Jenseits des Alphabets—Kommunikation mit Emojis". In: *Zeitschrift für germanistische Linguistik* 45.2, pp. 256–285.
- Gawne, Lauren and Gretchen McCulloch (2019). "Emoji as Digital Gestures". In: *language@internet* 17, Article 2.
- Grosz, Patrick G., Elsi Kaiser, and Francesco Pierini (2021). "Discourse anaphoricity and first-person indexicality in emoji resolution". In: *Proceedings of Sinn und Bedeutung* 25, pp. 340–357.
- Jackendoff, Ray and Eva Wittenberg (2014). "What you can say without syntax: A hierarchy of grammatical complexity". In: *Measuring linguistic complexity*. Ed. by Frederick J. Newmeyer and Laurel B. Preston. Oxford: Oxford University Press, pp. 65–82.
- Koch, Peter and Wulf Oesterreicher (1985). "Sprache der Nähe—Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte". In: *Romanistisches Jahrbuch* 36, pp. 15–43.
- Li, Li and Yue Yang (2018). "Pragmatic functions of emoji in internet-based communication—a corpus-based study". In: *Asian-Pacific Journal of Second and Foreign Language Education* 3, Article 16.
- McCulloch, Gretchen and Lauren Gawne (2018). "Emoji Grammar as Beat Gestures". In: *Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media (Emoji 2018)*.
- Pappert, Steffen (2017). "Zu kommunikativen Funktionen von Emojis in der WhatsApp-Kommunikation". In: *Empirische Erforschung internet-basierter Kommunikation*. Ed. by Michael Beißwenger. Berlin, Boston: De Gruyter, pp. 175–211.
- Pavalanathan, Umashanthi and Jacob Eisenstein (2016). "More emojis, less :) The competition for paralinguistic function in microblog writing". In: *First Monday* 21.11.
- Saussure, Ferdinand de (2011). *Grundfragen der allgemeinen Sprachwissenschaft*. Ed. by Charles Bally and Albert Sechehaye. [Original work published 1916]. Berlin, Boston: De Gruyter, pp. 70–75.
- Scheffler, Tatjana et al. (2022). "The processing of emoji-word substitutions: a self-paced-reading study". In: *Computers in Human Behavior* 127, p. 107076.
- Siebenhaar, Beat (2018). "Funktionen von Emojis und Altersabhängigkeit ihres Gebrauchs in der Whatsapp-Kommunikation". In: *Jugendsprachen/ Youth Languages: Aktuelle Perspektiven internationaler Forschung/Current Perspectives of International Research*. Ed. by Arne Ziegler. De Gruyter, pp. 749–772. URL: <https://doi.org/10.1515/9783110472226-034>.
- Weissman, Benjamin et al. (2023). "The lexicon of Emoji? Conventionality modulates processing of emoji". In: *Cognitive Science* 47.4, e13275. ISSN: 1551-6709. DOI: 10.1111/cogs.13275.

A. Demographic Information of Participants



B. Experiment 1: Items

ID	Syntax	Sentence	Group	SoSci
1	S-P-ADV	Die Sonne scheint warm.	A	A1
1i	ADV-P-S	Warm scheint die Sonne. 'The sun shines warmly.'	B	B1
2	S-P-ADV	Der Vogel singt laut.	B	B2
2i	ADV-P-S	Laut singt der Vogel. 'The bird sings loudly.'	A	A2
3	S-P-ADV	Das Auto fährt schnell.	A	A3
3i	ADV-P-S	Schnell fährt das Auto. 'The car drives fast.'	B	B3
4	S-P-ADV	Das Flugzeug fliegt hoch.	B	B4
4i	ADV-P-S	Hoch fliegt das Flugzeug. 'The plane flies high.'	A	A4
5	S-P-ADV	Die Ampel leuchtet grün.	A	A5
5i	ADV-P-S	Grün leuchtet die Ampel. 'The traffic light shows green.'	B	B5
6	S-P-O	Der Lehrer spielt Fußball.	B	B6
6i	O-P-S	Fußball spielt der Lehrer. 'The teacher plays soccer.'	A	A6
7	S-P-O	Die Königin isst Sushi.	A	A7
7i	O-P-S	Sushi isst die Königin. 'The queen eats sushi.'	B	B7
8	S-P-O	Der Weihnachtsmann liest ein Buch.	B	B8
8i	O-P-S	Ein Buch liest der Weihnachtsmann. 'Santa reads a book.'	A	A8
9	S-P-O	Die Hummel liebt Blumen.	A	A9
9i	O-P-S	Blumen liebt die Hummel. 'The bumblebee likes flowers.'	B	B9
10	S-P-O	Das Baby malt einen Regenbogen.	B	B10
10i	O-P-S	Einen Regenbogen malt das Baby. 'The baby paints a rainbow.'	A	A10

C. Experiment 2: Items

ID	Agent	Sentence	Grp	SoSci
11	human	Der Mann liegt im Bett.	A	A11
11t	animal	Der Hund liegt im Bett. 'The man/dog lies in bed.'	B	B11
12	human	Die Frau jongliert.	B	B12
12t	animal	Der Affe jongliert. 'The woman/monkey juggles.'	A	A12
13	human	Der Mann fährt Fahrrad.	A	A13
13t	animal	Das Eichhörnchen fährt Fahrrad. 'The man/squirrel rides a bicycle.'	B	B13
14	human	Die Frau fährt Ski.	B	B14
14t	animal	Der Hase fährt Ski. 'The woman/bunny skis.'	A	A14
15	human	Die Frau spielt Golf.	A	A15
15t	animal	Der Pinguin spielt Golf. 'The woman/penguin plays golf.'	B	B15
16	human	Peter schwimmt.	B	B16
16t	animal	Der Fisch schwimmt. 'Peter/The fish swims.'	A	A16
17	human	Leon und Florian halten Händchen.	A	A17
17t	animal	Die Otter halten Händchen. 'Leon and Florian/The otters hold hands.'	B	B17
18	human	Laura tanzt.	B	B18
18t	animal	Der Bär tanzt. 'Laura/The bear dances.'	A	A18
19	human	Mehmet spielt Basketball.	A	A19
19t	animal	Das Känguru spielt Basketball. 'Mehmet/The kangaroo plays basketball.'	B	B19
20	human	Aylin fährt Boot.	B	B20
20t	animal	Die Katze fährt Boot. 'Aylin/The cat rides in a boat.'	A	A20

Linear Featurality in Roman Script African Orthographies

David Roberts

Abstract. This paper uses the term “linear featurality” to describe the orthographic representation on the segmental (CV) tier of a unit lower on the phonological hierarchy than the phoneme. It is present, regularly or irregularly, in many of the world’s writing systems, and the Roman script orthographies of Africa are no exception. The paper surveys linear featurality in 180 languages from 30 countries, with language data extracted from published and unpublished literature, responses to an online forum, and a perusal of Omniglot, the online database of writing systems and languages. Cases identified include nasal vowels, prenasalization, labialization, palatalization, glottalized sounds, laterals, dentals, retroflexes, aspiration, rhoticization, fricativization, velarization, ATR, pharyngealization, back unroundedness, and clicks. The survey also illustrates some cases of simultaneous and irregular linear featurality, before a discussion comparing and contrasting linear and multilinear featurality in Africa.

I would like to thank Ronan Soni and Amalia Gnanadesikan for planting the seed for this article with the papers they presented on featurality at the *14th Workshop of the Association of Written Language & Literacy* in Rome in 2023 (Bezrukov & Soni 2023; Gnanadesikan 2023). My thanks are also due to the following people for providing language examples and/or contributing to discussions: Virginia Beavon-Ham, Carin Boone, Bethany Boston, Rod Casali, John Duerksen, Helen Eaton, Leoma Gilley, Amalia Gnanadesikan, Marian Hagg, Keir Hansford, Matthew Harley, Teresa Heath, Paul Humber, Andreas Joswig, Elke Karan, Eunice Kua, Robert Ladd, Jim Laesch, Susan Locklin, Constance Kutsch Lojenga, Jackie Marshall, Greg McLean, Pamela Morris, John Myhill, Russell Norton, Brian Parker, Rachel Penkoff, Robbie Petterson, Tim Pierce, Ron Radke, Ruth Raharimanantsoa, Tatiana Reid, JeDene Reeder, Bert Remijnsen, Paul Schaefer, Leila Schroeder, Keith Snider, Tim Stirtz, Wilma Wolthuis, Valentin Vydrine, and Lynell Zogbo. Thanks also to Ali Cripps for proofreading the chapter.

David Roberts  0009-0004-2282-9159
s/c BP57, Kara, Togo
E-mail: rbrdvd@gmail.com

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 95–121. <https://doi.org/10.36824/2024-graf-robe>
ISBN: 978-2-487055-10-0, e-ISBN: 978-2-487055-11-7

1. Introduction

Ever since Sampson (1985), it has been recognized that many writing systems contain featural elements, representing units lower on the phonological hierarchy than the phoneme. While most researchers agree that featurality does not warrant a separate typological class because no writing system is entirely and exclusively featural, it is increasingly acknowledged that featurality is widespread.

The newly emerging orthographies of African minority languages are no exception, yet featurality has received very little attention in the Africanist literature and is not generally taught in training programs for those engaged in orthography development. A twin article (Roberts in press) surveys *multilinear* featurality in Africa, i.e., the use of superscript and subscript diacritics to represent phonological features. The present article is a parallel survey focusing on *linear* featurality, that is, the weaving of featural elements into the tissue of the segmental tier. This tier is largely reserved for consonants and vowels, but the survey will also include the apostrophe <'> and saltillo <'> on the grounds that these symbols, unlike diacritics, occupy the same position as the vertical extensions in the letters <b d h k l>. Moreover, featurality researchers often dissect character stroke patterns, which would lead them to regard, for example, the upper hooks on the implosive letters <b d f g> as featural. This survey, in contrast, is limited to discrete whole characters, finding plenty to report on the featural use of these without adjusting the lens more finely.

The methodology involved a detailed perusal of published and unpublished sources in 180 languages, which inevitably means that some countries are better represented than others. Field workers in Cameroon and South Sudan have been particularly diligent in leaving an online paper trail. To help correct any geographical imbalance, I also accessed overviews such as *Omniglot*, the online encyclopedia of writing systems (Ager 1998–2024) and UNESCO's *Alphabets of Africa* (Hartell 1993). These, like many other individual sources, do not always clearly distinguish between existing practice and proposals, so the present survey makes no attempt to do so either. The survey also avoids reference to historical orthographies, although occasional mention will be made of recent reforms. Neither is there any attempt to assess the effectiveness in terms of literacy outcomes of the different featural strategies identified. The aim is more fundamental than this: to identify which phonological features are represented linearly, the degree of systematicity, and any areal tendencies.

This paper assumes that its intended readership is more likely to recognize phonetic labels (e.g., dental) than their more opaque counterparts (e.g., +CORONAL +ANTERIOR -DISTRIBUTED) as employed in Feature Geometry (Broe 1992; Clements 1985). This more flexible approach

also permits the survey to embrace phenomena such as rhotics, whose featural status is contested (Wiese 2011).

Orthographic data is presented between <chevrons> and phonemic data between /slashes/. The latter largely follow IPA, except that the palatal approximant is transcribed /y/ following the Africanist tradition. Transcription conventions in the sources have occasionally been slightly regularized for ease of comparison.

2. The Survey

2.1. Geographically Widespread Featurality

The survey begins by identifying five of the most frequent and widespread cases of linear featurality in Roman script African orthographies: nasal vowels (Section 2.1.1), prenasalized consonants (Section 2.1.2), labialized and palatalized consonants (Section 2.1.3), glottal consonants (Section 2.1.4), and lateral consonants (Section 2.1.5).

2.1.1. *Nasal Vowels*

The marking of nasal vowels with a postposed letter <n> is probably the most common and widespread case of linear featurality in African languages (Table 1, and see also Table 39). Much less frequently, postposed <ŋ> (Table 2) and <h> (Table 3) are also attested.

2.1.2. *Prenasalized Consonants*

The widespread strategy of representing prenasalized consonants with <m> preceding bilabial consonants and <n> elsewhere (Schroeder 2008, 30–31) is not strictly featural because -bilabial is not a natural class of sounds. However, many languages do represent prenasalization featurally, using <m> for labial consonants (Table 4), <n> for coronal consonants (Table 5), and <ŋ> for dorsal consonants (Table 6, and see also Tables 36, 38, and 42). Depending on the language, labio-velars participate in either the labial or dorsal graphic series, and labio-dentals in either the labial or coronal graphic series, either choice being justifiable from a featural point of view. Sources do not always clearly distinguish between prenasalized consonants /^mC/ and preconsonantal syllabic nasals /ŋC/; Tables 4–6 present them as units for the sake of regularity. Either way, the orthographic representations are featural.

Another featural strategy is to represent all prenasalized consonants, irrespective of their place of articulation, with a preposed <n>. In such cases, only the manner, not the place of articulation, is specified featurally (Table 7).

TABLE 1. Representation of nasal vowels with postposed <n>

Speech	Writing	Language	Family	Country	Citation
/Ṽ/	<Vn>	Aja	Kwa	Benin	(CISOLA 2010, 8-9)
		Glui	Kx'a	Botswana	(Omniglot)
		Cerma	Gur	Burkina Faso	(Hartell 1993, 42)
		Tikar	Bantoid	Cameroon	(Hartell 1993, 93)
		Nzakara	Ubangi	Central African Republic	(ACATBA 2005, 48)
		Niaboua	Kru	Côte d'Ivoire	(Hartell 1993, 140, 165, 194)
		Nafaanra	Gur	Ghana	(Drame & Hagg 2020, 11)
		Grebo	Kru	Liberia	
		Soninke	Mande	Mali	(Legbeti 2016, 89)
		Ososo	Edoid	Nigeria	(Hartell 1993, 282)
		Anufɔ	Kwa	Togo	

TABLE 2. Representation of nasal vowels with postposed <ŋ>

/Ṽ/	<Vŋ>	Gwama	Nilo-Saharan	Ethiopia	(Hellenthal & Kutsch Lojenga 2011, 8)
		Lama	Gur	Togo	(Hartell 1993, 289)

TABLE 3. Representation of nasal vowels with postposed <h>

/Ṽ/	<Vh>	Mo Vagala	Gur	Ghana	(Hartell 1993, 164, 167)
-----	------	-----------	-----	-------	--------------------------

2.1.3. *Labialized and palatalized consonants*

Labialization and palatalization are extremely common in African languages, and are especially prolific in Cameroon and Nigeria. These features are usually represented with the rather obvious strategy of postposed <w y> respectively (Table 8). See also Section 2.3, Tables 36–38. Many other African languages have only labialization, but I am unaware of any that only have palatalization.

Alternative featural strategies, which to my knowledge are confined to Cameroon and South Sudan, represent labialization and palataliza-

TABLE 4. Representation of prenasalized labial consonants with preposed <m>

/ ^m b ^{m̃} g̃b/	<mb mgb	Isu	Grassfields	Cameroon	(Anderson 2014, 5)
^{m̃} f ^{m̃} pf	mf mpf				
^{m̃} bv ^{m̃} m/	mbv mm>				
/ ^{m̃} b ^{m̃} v/	<mb mv>	Budu	Bantu	DR Congo	(Bamata 1997, 2)
/ ^{m̃} b ^{m̃} f	<mb mf	Yilumbu		Gabon	(Hubert & Mavoungou 2010, 183)
^{m̃} v/	mv>				
/ ^{m̃} p ^{m̃} b	<mp mb	Mbembe	Cross River	Nigeria	(Barnwell ms, 5)
^{m̃} k̃p	mkp				
^{m̃} g̃b ^{m̃} φ	mgb mph				
^{m̃} β ^{m̃} f ^{m̃} v	mbh mf mv				
^{m̃} m/	mm>				
/ ^{m̃} p ^{m̃} b/	<mp mb>	Ngando	Bantu	Central African Republic	(Murrell 2020, 4-5)
/ ^{m̃} b ^{m̃} g̃b	<mb mgb	Ma'di	C Sudanic	S Sudan	(Kilpatrick 2004, 89-90)
^{m̃} v/	mv>				
/ ^{m̃} p ^{m̃} b/	<mp mb>	Gusilay	Atlantic	Senegal	(Diatla 2013, 3, 7)
/ ^{m̃} b ^{m̃} g̃b	<mb mgb	Lugbariti	C Sudanic	Uganda	(Anonymous ms, 5)
^{m̃} v/	mv>				

TABLE 5. Representation of prenasalized coronal consonants with preposed <n>

/ ^{ñ} d ^{ñ} s ^{ñ} z	<nd ns nz	Bakoko	Bantu	Cameroon	(Njeck & Anderson 2010, 4)
^{ñ} d̃ʒ ^{ñ} l̃ ^{ñ} y/	nj nl ny>				(Prinz ms, 3)
/ ^{ñ} d ^{ñ} d̃ʒ/	<nd nj>	Massalit	Nilo-Saharan Maban	Chad	(Hubert & Mavoungou 2010, 183)
/ ^{ñ} d ^{ñ} z	<nd nz	Yilumbu	Bantu	Gabon	(Hall & Hall 2004, 81-82)
^{ñ} ʃs ^{ñ} d̃ʒ/	nts ndz>				
/ ^{ñ} d ^{ñ} z ^{ñ} ʃ/	<nd nz nj>	Jur Modo	C Sudanic	S Sudan	

TABLE 6. Representation of prenasalized dorsal consonants with preposed <ŋ>

/ ^{ŋ̃} k ^{ŋ̃} g	<ŋk ŋg	Mungong	Beboid	Cameroon	(Mkounga 2022, 5)
^{ŋ̃} g̃b ^{ŋ̃} y/	ŋgb ŋgh>				
/ ^{ŋ̃} g ^{ŋ̃} m̃g̃b/	<ŋg ŋb>	Jur Modo	C Sudanic	S Sudan	(Hall & Hall 2004, 81)

TABLE 7. Representation of all prenasalized consonants with preposed <n>

/ ^{m̃} b ^{ñ} d ^{ŋ̃} g/	<nb nd ng>	Kisie	Atlantic	Guinea	(Hartell 1993, 174)
/ ^{ñ} t ^{ñ} c ^{ŋ̃} k	<nt nc nk	Karon		Senegal	(SIL 2011, 18)
^{m̃} f ^{ñ} s ^{ŋ̃} h/	nf ns nh>				

tion with postposed vowels <Cu Ci> respectively (Table 9) or, in several Beboïd languages of Cameroon, asymmetrically as <Cw Ci> (Table 10).

Finally, some Grassfields languages of Cameroon represent a third series of labio-palatalized consonants featurally (Table 11), and at least one Plateau language of Nigeria does so for lengthened labialization and palatalization (Table 12).

2.1.4. *Glottal Sounds*

Implosive consonants are common in many parts of Africa, although the palatal /f/ and velar /g/ positions are much less frequently attested than the labial /b/ and alveolar /d/ positions (Sande & Oakley 2023, 2). Phonemic transcriptions in the literature tend to vary between the palatal implosive /f/ and the glottalised palatal approximant /ʔy/. Implosives form a natural class with other glottal sounds, and tend to be represented featurally either with a preposed apostrophe (Table 13), or a postposed letter <h> (Table 14, and see also Tables 32 and 42).

Ejective consonants are abundant in Ethiopia and also occur sporadically elsewhere. Like implosives, they form a natural class with other glottal sounds, and are usually represented featurally with a postposed apostrophe <'> (Table 15). The use of postposed <h> in Blin (Table 43) is exceptional. At least one Ethiopian language uses separate featural strategies for representing implosives and ejectives (Table 16). However, since implosives and ejectives both involve glottal release and never co-occur (the former being voiced and the latter voiceless), it is completely defensible, from a featural point of view, to represent them with a single orthographic strategy such as a postposed apostrophe <'> (Table 17) or a preposed apostrophe <'> (Table 18).

2.1.5. *Lateral Consonants*

In a wide geographical spread of languages, with a concentration in Chadic, lateral consonants are marked featurally with the letter <l> (Table 19).

2.2. Geographically restricted featurality

The remaining cases of linear featurality in African Roman Script orthographies are much less frequently attested and geographically more restricted. Beginning with consonants, the survey will cover six kinds: dental consonants (Section 2.2.1), retroflex consonants (Section 2.2.2), aspiration (Section 2.2.3), rhoticization (Section 2.2.4), and fricativization (Section 2.2.5). Then, turning to vowels, three cases will be enumer-

TABLE 8. Representation of labialization and palatalization with postposed <w y> respectively

/C ^w Cy/ <Cw Cy>	Tɔfɛn S Toussian Akoose	Kwa Gur Bantu	Benin Burkina Faso Cameroon	(Hartell 1993, 34) (SIL 2001, 5) (Hedinger 2011, 9-10)
	Moghamo Ipulo Luumbu Senara	Grassfields Tivoid Bantu Senoufo		(Njeck 2008, 4-5) (Tuinstra 2018, 9-12) (Mfouhou 2022, 8) (Hartell 1993, 141, 300)
	Fuliru Latege Tharaka Bwaatye	Bantu Chadic	DR Congo Gabon Kenya Nigeria	(Linton 2016, 4) (Schroeder 2008, 31) (Pweddson n.d., 122-125)
	Bekwarra Tiv	Cross River Tivoid		(Hartell 1993, 221) (Iornumbe 2020, 51-55)
	Venda	Bantu	S Africa	(PanSALB 2019, 10-14)
	Zinza Ruuli-Nyala		Tanzania Uganda	(Odom 2015, 7) (Nzogi 2011, 5-6)

TABLE 9. Representation of labialization and palatalization with postposed <u i> respectively

/C ^w Cy/ <Cu Ci>	Nulibie Numala Nuasue Otuho Dinka Nuer	Bantu Nilotic	Cameroon S Sudan	(Boyd 2015a, 6-8) (Boyd 2015b, 4-7) (Boyd 2016, 4-6) (Duerksen 2004b, 40, 71)
-----------------------------	---	----------------------	-------------------------	--

TABLE 10. Representation of labialization and palatalization with postposed <w i> respectively

/C ^w Cy/ <Cw Ci>	Kemezong Nooni Mungong Naami	Beboid	Cameroon	(Cox 2005, 4-5) (Lux 2018, 7-9) (Mkouna 2022, 6-8) (Mkouna 2023, 12-14)
-----------------------------	---------------------------------------	--------	----------	--

TABLE 11. Representation of labio-palatalization with postposed <wy>

/C ^w C ^y C ^q /	<C ^w C ^y C ^{wy} >	Mfumte	Grassfields	Cameroon	(Eyoh & Hedinger 2008, 5)
		Koffa			(McLean 2020a, 5-7)
		Ndə Wuli			(McLean 2020d, 6-7)
		Kwaja			(McLean 2020b, 5-6)

TABLE 12. Representation of lengthened labialization and palatalization with postposed <ww yy>

/C ^w C ^{ww} C ^y C ^{yy} /	<C ^w C ^{ww} Gworok C ^y C ^{yy} >	Plateau	Nigeria	(Anonymous 2015, 13-18)
---	--	---------	---------	----------------------------

ated: ATR (Section 2.2.6), pharyngealization (Section 2.2.7), and back unroundedness (Section 2.2.8).

2.2.1. *Dental Consonants*

A concentration of languages in Sudan and South Sudan contrasts dental and alveolar plosives, with the dental series usually being represented featurally with a postposed letter <h> (Table 20, and see also Table 35). However, at least two languages use the same strategy to mark the non-dental series (Table 21).

2.2.2. *Retroflex Consonants*

At least one Ethiopian language marks a series of retroflex consonants with a postposed apostrophe <'> (Table 22).

2.2.3. *Aspirated Consonants*

Aspirated consonants sometimes contrast with their non-aspirated counterparts, and the aspirated series is sometimes represented featurally with a postposed <h> (Table 23, and see also Tables 32 and 33).

2.2.4. *Rhoticized Consonants*

The inventories of some languages in Cameroon and South Sudan include a series of rhoticized consonants that are invariably represented with a postposed <r> (Table 24, and see also Tables 38 and 42). Sources do not always make it clear whether /Cr/ is analyzed as a sequence or a unit, but Table 24 presents them as units for the sake of regularity. Either way, the orthographic representations are featural.

TABLE 13. Representation of implosives and other glottal sounds with preposed apostrophe <'>

/b d/	<'b 'd>	Lele Mono	Chadic Ubangi	Chad DR Congo	(Hartell 1993, 113)
/b ʔ ʔw ʔy/	<'b 'l 'w 'y>	S Birifor	Gur	Ghana	(Omniglot)
/b d ʔw ʔy/	<'b 'd 'w 'y>	Aringa	C Sudanic	S Sudan	(Kilpatrick 2004, 90)
/b d f/	<'b 'd 'j>	Ma'diti Beli			(SIL ms-b, 44)
	<'b 'd 'y>	Baka Bari	Nilotic		(Stirtz 2014, 26)
/b d/	<'b 'd>	Mundu Kadugli	Ubangi Kadu	Sudan	(Persson 2004, 81)
					(SIL ms-b, 13)
/p ʔ ʔw ʔy/	<'p 'l 'w 'y>	Teen	Gur	Côte d'Ivoire	(Bilal 2004, 182)
					(Hall & Hall 2004, 59)
					(Hartell 1993, 142)

TABLE 14. Representation of implosives and other glottal sounds with postposed <h>

/b d f/	<bh dh vh>	Mangbetu	C Sudanic	DR Congo	(Hartell 1993, 303)
/b d/	<bh dh>	E Dan (2014)	Mande	Côte d'Ivoire	(Vydrin et al. 2019, 9)
		Ngando	Bantu	Central African Republic	(Murrell 2020, 3)
		Lubwisi	Bantu	Uganda	(Tabb et al. 2011, 10, 14)
/aʔ eʔ oʔ/	<ah eh oh>	S Toussian	Gur	Burkina Faso	(SIL 2001, 6-7)

TABLE 15. Representation of ejectives with postposed apostrophe <'>

/k ʔs' ʔj'/	<k' s' c'>	Mpadi	Chadic	Cameroon	(Allison 2006, 2-4)
/t' k'/	<t' k'>	Gawwada	Cushitic	Ethiopia	(Omniglot)
/p' t' k' ts'/	<p' t' k' c'>	Bambassi	Omotic		
/p' t' ts' tʂ' ʔj'/	<p' t' x' c' ch'>	Bench			
/p' k' ʔj'/	<p' k' c'>	Kafa			
/p' t' k' ʔj'/	<p' t' k' c'>	Waata	Cushitic	Kenya	

TABLE 16. Separate representation of implosives with postposed <h> and ejectives with postposed apostrophe <'>

Implosives	/b̥ d̥ g̥/	<bh dh gh>	Hamar	Omoti	Ethiopia	(Yigezu & Mendisu 2015, 12-14)
Ejectives	/s' q' tʃ'/	<s' q' c'>				

TABLE 17. Unified representation of ejectives and implosives with postposed apostrophe <'>

/p' t' k' f̥s' tʃ' d̥/	<p' t' k' s' c' d'>	Borna	Omoti	Ethiopia	(Joswig et al. 2010, 2-3)
/p' t' c' k' f̥s' tʃ' b̥ d̥/	<p' t' c' k' s' ç' b' d'>	Gumuz	Nilo-Saharan satellite		(Omniglot)

TABLE 18. Unified representation of ejectives and implosives with a preposed apostrophe <'>

/p' t' t̥ c' k' b̥ d̥/	<p' t' t̥ c' k' b' d'>	Uduk	Nilo-Saharan satellite	S Sudan	(Omniglot)
------------------------	------------------------	------	------------------------	---------	------------

TABLE 19. Representation of lateral consonants with <l>

/l tɬ/	<l tl>	Tswana	Bantu	Botswana	(Omniglot)
/l ɬ ɬ̥/	<l sl zl>	Buwal	Chadic	Cameroon	(Viljoen et al. 2009, 3-5)
		Guemzek			(Scherrer 2014, 8)
		Muyang			(Smith 2001, 5-6)
		Vame			(Kinnaid 2005, 5-6)
	<l tl dl>	Musey Ngizim		Chad Nigeria	(Omniglot)
/l ɬ ɬ̥ kɬ/	<l hl dl kl>	Swati	Bantu	Eswatini	
/l ɬ ɬ̥ t̥/	<l hl dl tl>	Tswa		Mozambique	
/l p̥ b̥ t̥ k̥ p̥/	<l pl bl tl kpl>	Idoma	Idomoid	Nigeria	(Armstrong ms, 8-9)
f̥ m̥ n̥ w̥/	fl ml nl wl>				(Omniglot)
/l ɬ ɬ̥ kɬ t̥/	<l hl dl kl tl>	S Ndebele	Bantu	S Africa	
/ɬ t̥/	<sl tl>	Iraqw	Cushitic	Tanzania	

2.2.5. Fricativized Consonants

At least one Bantu language of Mozambique marks fricativized consonants featurally with a postposed <v> (Table 25), and at least one Cameroonian language has a series of consonants with velar fricative

TABLE 20. Representation of dental plosives with postposed <h>

/t̥ d̥ n̥/	<th dh ndh>	Olu'bo	C Sudanic	S Sudan	(Kilpatrick 2004, 89)
/t̥ d̥ n̥/	<th dh nh>	Anywa	Nilotic		(Gilley 2004, 173)
		Shilluk			
		Dinka			(Duerksen 2004a, 74)
		Nuer			
/d̥ n̥/	<dh nh>	Belanda Boor			(Gilley 2004, 173)

TABLE 21. Representation of non-dental plosives with postposed <h>

/t d/	<th dh>	Jumjum	Nilotic	S Sudan	(Gilley 2004, 173)
		Burun			

TABLE 22. Representation of retroflex consonants with postposed apostrophe <'>

/tʂ ʂ ʒ tʂ' /	<c' s' z' x'>	Sheko	Omoti	Ethiopia	(Omniglot)
---------------	---------------	-------	-------	----------	------------

TABLE 23. Representation of aspiration with postposed <h>

/p ^h t ^h k ^h /	<ph th kh>	Kaansa	Gur	Burkina Faso	(SIL ms-a, 16, 20)
---	------------	--------	-----	--------------	--------------------

TABLE 24. Representation of rhotics with postposed <r>

/tʳ fʳ lʳ/	<tr fr dr>	Baba 1	Grassfields	Cameroon	(Njeck & Hedinger 2006, 5)
/pʳ tʳ kʳ fʳ	<pr tr kr fr>	Bambalang			(Grove & Wright 2012, 11)
ʃʳ ɣʳ ʈʳ lʳ	shr ghr chr lr				
yʳ wʳ/	yr wr>				
/tʳ dʳ ndʳ /	<tr dr ndr>	Avokaya	C Sudanic	S Sudan	(Kilpatrick 2004, 89-90)
/tʳ kʳ dʳ	<tr kpr dr>	Baka			(Persson 2004, 82)
ɡʳ bʳ ndʳ/	gbr ndr>				
/pʳ tʳ bʳ dʳ/	<pr tr br dr>	Morokodo			

TABLE 25. Representation of fricativized consonants with postposed <v>

/bʷ sʷ zʷ/	<bv sv zv>	Nyungwe	Bantu	Mozambique	(Omniglot)
------------	------------	---------	-------	------------	------------

TABLE 26. Representation of velar fricative articulation with postposed <gh>

/ɣ ɳɣ bɣ fɣ/	<gh ngh bgh fgh>	Weh	Grassfields	Cameroon	(Davison 2009, 8)
--------------	------------------	-----	-------------	----------	-------------------

primary or secondary articulation. These are marked featurally with the digraph <gh> (Table 26).

2.2.6. *ATR Vowels*

Certain Sudanese languages mark ATR with a postposed <x> (Table 27, and see also Table 46). However, some have recently rejected this linear strategy in favor of multilinear superscript umlauts (Alaki & Norton 2015, 228). A postposed <h> is also attested in one Nigerian language (Table 28, and see also Table 34). A few sources advocate marking the orthographic word with only one symbol, for languages in which all the vowels in a phonological word agree in ATR quality (Table 29).

TABLE 27. Representation of +ATR vowels with postposed <x>

/i e ə o u/	<ix ex ax ox ux>	Kamda	Kadu	Sudan	(Hall & Hall 2004, 62)
		These	E Sudanic		(Yip 2004, 116)

TABLE 28. Representation of +ATR vowels with postposed <h>

/e o/	<eh oh>	Dhuo	Mambiloid	Nigeria	(Hartell 1993, 235)
-------	---------	------	-----------	---------	---------------------

TABLE 29. Representation of +ATR vowels with word final <h>

+ATR vowels	word final <h>	Ogba Igede	Igboid Idomoid	Nigeria	(Williamson 1984, 24-25)
-------------	----------------	---------------	-------------------	---------	-----------------------------

Hartell (1993, 109) reports a highly unusual application of the ATR word-marking principle in Zande (Ubangi, Central African Republic), where the letter <j> is placed “after the first simple (double) vowel of the word” to indicate that the whole word is +ATR.

2.2.7. *Pharyngealized Vowels*

Several Kx’a languages of southern Africa have pharyngealized vowels, which are represented featurally with a postposed <q> (Table 30, and see also Table 39).

TABLE 30. Representation of pharyngealized vowels with postposed <q>

/ɑ̣ ɛ̣ ɪ̣ ɔ̣/ ʊ̣/	<aq eq iq oq uq>	ǀ'Amkoe	Kx'a	Botswana	(Omniglot)
/ɑ̣ ɔ̣ ʊ̣/	<aq oq uq>	Ekoka !Kung		Namibia	
/ɑ̣ ɔ̣/	<aq oq>	Ju 'hoan			

2.2.8. Back Unrounded Vowels

One Togolese language uses the postposed letter <ɣ> featurally to unify a series of five long back unrounded vowels (Table 31).

TABLE 31. Representation of long back unround vowels with postposed <ɣ>

/uɣ ɯɣ ɤɣ ʌɣ ɑɣ/	<iɣ ɪɣ eɣ ɛɣ aɣ>	Kabiye	Gur	Togo	(Roberts 2008, 54-55)
------------------	------------------	--------	-----	------	-----------------------

2.3. Simultaneous featurality

This section illustrates some cases of simultaneous featurality, drawing attention to two possible scenarios. First, a single orthographic convention may represent more than one feature, and within each series, do so regularly and completely. In Ekoka !Kung, the postposed <h> simultaneously represents aspiration, glottalization, and breathiness (Table 32).

TABLE 32. Simultaneous representation of aspiration, glottalization, and breathiness with postposed <h>

Aspiration	/p ^h t ^h k ^h tʃ ^h /	<ph th kh tch>	Ekoka !Kung	Kx'a	Namibia (Omniglot)
Glottalization	/b ^h d ^h g ^h m ^h n ^h /	<bh dh gh mh nh>			
Breathiness	/ə̤ e̤ i̤ ʊ̤ ṳ/	<ah eh ih oh uh>			

Nyungwe uses the same strategy to simultaneously represent voiceless aspirated (as opposed to non-aspirated) consonants, and voiced plosive (as opposed to implosive) consonants (Table 33).

Luwo uses the same strategy to simultaneously represent both dental consonants and +ATR vowels (Table 34).

TABLE 33. Simultaneous representation of aspiration and voiced plosives with postposed <h>

Aspiration	/p ^h t ^h k ^h /	<ph th kh>	Nyungwe Bantu Mozambique (Omniglot)
Voiced plosives	<b d>	<bh dh>	

TABLE 34. Simultaneous representation of dental consonants and +ATR vowels with postposed <h>

Dentals	/t̪ d̪ n̪/	<th dh nh>	Luwo	Nilotic	S Sudan	(Gilley 2004, 173)
+ATR	/ə e i o u/	/ah eh ih oh uh/				

Finally, Goemai uses the postposed apostrophe <'> to simultaneously represent both voiceless non-aspirated (as opposed to aspirated) consonants, and implosive (as opposed to voiced plosive) consonants (Table 35).

TABLE 35. Simultaneous representation of non-aspiration and implosives with postposed apostrophe <'>

Non-aspiration	/t k f s ʃ/	<t' k' f' s' sh'>	Goemai	Chadic	Nigeria	(Omniglot)
Implosives	/ɓ d/	<b' d'>				

The second scenario is equally widespread. A single series of graphemes may systematically manifest more than one featural characteristic. Labialization is probably the articulation most likely to participate in this. Numerous Cameroonian languages combine it with labial, coronal, and/or dorsal prenasalization (<NCw>, Table 36a, b, c respectively), while elsewhere labialization co-occurs with laterals <Clw> and rhotics <Crw> (Table 37a, b respectively).

Convergences also frequently occur between prenasalization and another feature, such as palatalization <NCy> or rhotics <NCr> (Table 38a and b respectively).

Similarly, in some Kx'a languages, nasalization co-occurs with pharyngealization <Vqn> or breathiness <Vhn> (Table 39a and b respectively).

Finally, these Kx'a languages, as well as Khoe-Kwadi and some of Bantu stock, are characterized by an elaborate series of clicks with multiple secondary and tertiary articulations, which are represented orthographically with extensive and often highly regular simultane-

TABLE 36. Simultaneous representation of labialization with (a) labial, (b) coronal, and (c) dorsal prenasalization

a	$/\text{m}^{\text{pw}} \text{mb}^{\text{w}} \text{nf}^{\text{w}}/$	<mpw mbw mfw>	Sari	Beboid	Cameroon	(Langhout 2022, 9-10)
b	$/\text{nt}^{\text{w}} \text{nd}^{\text{z}} \text{nd}^{\text{w}} \text{nf}^{\text{w}}/$	<ntw nzw ndw nchw>	Bafanji	Grassfields		(Hamm & Hamm 2016, 8)
c	$/\text{ng}^{\text{w}} \text{ng}^{\text{m}} \text{gb}^{\text{w}} \text{nv}^{\text{w}}/$	<ngw ngvw ngvw>	Kwanja	Mambiloid		(Weber & Weber 2008, 5)

TABLE 37. Simultaneous representation of labialization with (a) laterals and (b) rhotics

a	$/\text{tl}^{\text{lw}} \text{tl}^{\text{wh}} \text{dl}^{\text{w}}/$	<tlw tlhw dlw>	Tsonga	Bantu	S Africa	(Omniglot)
b	$/\text{tr}^{\text{rw}} \text{dr}^{\text{w}} \text{nd}^{\text{rw}}/$	<trw drw ndrww>	Moru	C Sudanic	S Sudan	(Kilpatrick 2004, 90)

TABLE 38. Simultaneous representation of prenasalization with (a) palatalization and (b) rhotics

a	$/\text{nt}^{\text{y}} \text{nt}^{\text{f}} \text{nd}^{\text{z}} \text{nd}^{\text{y}} \text{nd}^{\text{f}} \text{mb}^{\text{y}} \text{mf}^{\text{y}}/$	<nty nchy nzy ndy> <mpi mbi mfi>	Bafanji	Grassfields	Cameroon	(Hamm & Hamm 2016, 6) (Langhout 2022, 8)
b	$/\text{nt}^{\text{r}} \text{nd}^{\text{r}} \text{ns}^{\text{r}} \text{nz}^{\text{r}} \text{nf}^{\text{r}}/$	<ntr ndr nsr nzs nrhr>	Mbembe	Jukunoid		(Eyoh & Sandamu 2009, 8)

TABLE 39. Simultaneous representation of nasalization with (a) pharyngealization and (b) breathiness

a	$/\tilde{\text{a}}^{\text{f}} \tilde{\text{o}}^{\text{f}} \tilde{\text{a}}^{\text{f}} \tilde{\text{o}}^{\text{f}} \tilde{\text{a}}^{\text{f}} \tilde{\text{o}}^{\text{f}}/$	<aqn oqn> <aqn oqn uqn>	Ju 'hoan	Kx'a	Namibia	(Omniglot)
b	$/\tilde{\text{a}} \tilde{\text{o}}/$	<ahn ohn>	Ekoka !Kung			

ous featureality. The five ʘ-Amkoe clicks—labial, dental, lateral, alveolar, palatal—are represented featurally with the IPA symbols <ʘ | ǀ | ǁ | ǃ> respectively, while a host of secondary articulations are also represented, such as uvulars with the postposed letter <q> (Table 40).

Table 41 summarizes how the clicks themselves are represented orthographically in other southern African languages, while Table 42 summarizes how a representative sample of secondary articulations is represented.

TABLE 40. Simultaneous representation of clicks and uvular secondary articulations in ǀʼAmkoe (Botswana)

	Bilabial	Dental	Lateral	Alveolar	Palatal
Simple uvular	<Ōq>	< q>	< q>	<!q>	<ǀq>
Ejective uvular	<Ōqʼ>	< qʼ>	< qʼ>	<!qʼ>	<ǀqʼ>
Aspirated uvular	<Ōqh>	< qh>	< qh>	<!qh>	<ǀqh>
Voiced ejective uvular		<g qʼ>	<g qʼ>		
Ejective affricate uvular	<Ōqx>	< qx>	< qx>	<!qx>	<ǀqx>
Voiced ejective affricate uvular	<gŌqx>		<g qx>		<gǀqx>

TABLE 41. Symbolization of clicks in other southern African languages

Dental	/ /	< >	G ui Ekoka !Kung Ju 'hoan Khoekhoe	Kx'a Khoe-Kwadi	Botswana Namibia	(Omniglot)
		<c>	Naro Swati N Ndebele S Ndebele Zulu	Bantu	Botswana Eswatini S Africa	
		<q>	Tsonga			
Lateral	/ /	< >	G ui Ekoka !Kung Ju 'hoan Khoekhoe	Kx'a Khoe-Kwadi	Botswana Namibia	
		<x>	Naro S Ndebele N Ndebele	Bantu	Botswana S Africa	
		<q>	Zulu			
Alveolar	/!/	<!>	G ui Ekoka !Kung Ju 'hoan Khoekhoe	Kx'a Khoe-Kwadi	Botswana Namibia	
		<q>	Naro N Ndebele S Ndebele Zulu	Bantu	Botswana S Africa	
Palatal	ǀ	<ǀ>	G ui Khoekhoe Ekoka !Kung Ju 'hoan	Kx'a Khoe-Kwadi Kx'a	Botswana Namibia	
		<x>	Zulu	Bantu	S Africa	

TABLE 42. Representative sample of secondary articulation symbolization in southern African click languages

Aspirated	/C ^h / <Ch>	G ui Naro ǀ'Amkoe Khoekhoe Ekoka !Kung Ju 'hoan	Khoe-Kwadi Kx'a Khoe-Kwadi Kx'a	Botswana (Omniglot) Namibia
Glottalized	/Cʔ/ <C'>	Naro ǀ'Amkoe Ju 'hoan	Khoe-Kwadi Kx'a	Botswana Namibia
Nasalized	/ᵑC/ <nC>	G ui Naro ǀ'Amkoe Ekoka !Kung Ju 'hoan S Ndebele N Ndebele	Khoe-Kwadi Kx'a	Botswana Namibia
Voiced	/gC/ <gC>	G ui ǀ'Amkoe Ekoka !Kung Ju 'hoan S Ndebele Zulu	Khoe-Kwadi Kx'a	Botswana Namibia S Africa

2.4. Irregular featurality

The survey so far may have given the impression of a high degree of systematic featurality in African orthographies, whereas in fact irregularity is just as widespread. Highlighting the following cases carries no implication that the strategies are necessarily defective, since decision-makers may be influenced by linguistic or sociolinguistic issues that go unexplained in the sources consulted. Again, it will be helpful to highlight two possible strategies. In all the following examples, the regular featurality is highlighted in bold.

The most common scenario is when a single orthographic strategy is applied regularly within a particular featural series, but is also employed irregularly for a mixed bag of other unrelated purposes. Table 43 illustrates this for implosives, ejectives, dentals, retroflexes, aspiration, and velarization in various languages.

The second scenario, when featurality is applied partially but does not extend to an entire series, is also widespread. In languages that mark fricatives and affricates with a postposed <h>, the phoneme /d͡ʒ/, if present, is usually excluded, being represented non-featurally as <j> (Table 44). To a lesser extent, the same is true of /t͡ʃ/, which is sometimes represented non-featurally as <c> (Table 45). I am unaware of any

African language that marks the fricative/affricate series with entirely regular featurality.

Similarly, three languages of South Sudan mark all but one of the series of +ATR vowels with a postposed <x> (Table 46).

Finally, in Tswana, the irregularity operates in both directions. On the one hand, the postposed <h> falls one short of marking all the aspirated consonants and, on the other hand, it marks one unaspirated consonant (Table 47).

3. Discussion

Featurality bequeaths visual similarity to an orthography, binding together a natural class of sounds. As the foregoing survey has amply demonstrated, linear featurality, which weaves featural elements into the tissue of the segmental tier, is widespread in Roman script African orthographies.

Table 48 summarizes symbol choice and position for the linear featural representations of vowels. In all cases, they involve digraphs the second element of which is a letter, most commonly <Vh>, which may represent nasality, ATR, or breathiness. Of the three options for nasal vowels, the digraph <Vn> is by far the most frequently attested and geographically widespread.

Click consonants tend to be represented with their corresponding IPA characters, but the letters <c q x> are also attested (Table 49). Obviously, none of these orthographic strategies is geographically widespread because the clicks themselves are not.

The picture that emerges for consonants other than clicks is more complex because they are represented with an additional symbol which may be (a) preposed or postposed, (b) a consonant, a vowel, or an apostrophe, and these form (c) digraphs or, in one case, a trigraph (Table 50).

With these summary charts in mind, it will be instructive, by way of conclusion, to compare the two basic approaches to featurality—linear and multilinear—identifying which characteristics they share and to what extent they differ.

Linear and multilinear strategies encode a similar number of phonological features. Nasal vowels, which are ubiquitous, attract linear and multilinear featurality in more or less equal measure (<Vn> vs <Ṽ>; Roberts in press, xx), and the same is true of implosives (<bh dh...> vs <b̥ d̥...>; Roberts in press, xx). Tongue-root position is another matter. In areas of Africa that are resistant to representing the -ATR set with special characters, multilinear featurality is clearly the preferred option, albeit with a startling lack of uniformity (Roberts in press, xx). The linear strategies reported in this paper, on the other hand, are much less

TABLE 43. Extending a featural strategy irregularly for other unrelated purposes

Implosives	/b d f ɸ ɹ kʷ/ ɸ x/ θ ð nð c/ t d f ʒ/ pʰ kʰ tʃʰ ʃ ɣ/ kʷ bʷ gʷ bvʷ ʃ ɣ tʃ/	<bh dh sh ch zh wh> <th qh ch sh kh> <th dh ndh ch> <th dh sh zh> <ph kh ch sh gh> <kh bh gh bvʰ sh gh ch>	Ekpeye Blin Dholuo Kanga Tumbuka Isu	Igboide Cushitic Nilotic Kadu Bantu Grassfields	Nigeria Eritrea Kenya Sudan Malawi Cameroon	(Ngulube 2011a, 56) (Fallon 2006, 96) (Hartell 1993, 183) (Hall & Hall 2004, 59) (Omniglot) (Anderson 2014, 4, 8)
------------	---	--	---	--	--	---

TABLE 44. Omitting /dʒ/ in a series of fricatives and affricates

/ʃ ʒ ʧ tʃ dʒ/	<sh zh gh ch j>	Lidzonka Ndə Wuli	Grassfields	Cameroon	(McLean 2020c, 4-5) (McLean 2020d, 3-4)
/ʃ ʧ tʃ dʒ/	<sh gh ch j>	Awing			(Alomofor & Anderson 2005, 2)
/ʃ tʃ dʒ/	<sh ch zh j>	Mbembe	Jukunoid		(Eyoh & Sandamu 2009, 5-6)
/ʃ ʒ x ʧ tʃ dʒ/	<sh zh kh gh ch j>	Tera	Chadic	Nigeria	(Tench n.d., 19)
/ʃ ʒ tʃ dʒ/	<sh zh ch j>	Lunda	Bantu	Zambia	(Omniglot)

TABLE 45. Omitting /tʃ dʒ/ in a series of fricatives and affricates

/ʃ ʒ tʃ dʒ/	<sh zh c j>	Mekaa	Bantu	Cameroon	(Hartell 1993, 73, 84)
/ʃ ʒ ʧ tʃ dʒ / /ʃ ʒ tʃ dʒ/	<sh zh gh c j> <sh zh c j>	Fe'efe'e Nupe	Grassfields Nupoid	Nigeria	(Madugu n.d., 32, 35-36)
/ʃ ʒ ʧ tʃ dz/	<sh zh gh c j>	Tarok	Plateau		(Hartell 1993, 239)

TABLE 46. Omitting /ə/ in a series of +ATR vowels

/i e o u ə/	<ix ex ox ux a>	Kadugli Kanga Kufa	Kadu	Sudan	(Hall & Hall 2004, 62-63)
-------------	-----------------	--------------------------	------	-------	------------------------------

TABLE 47. Irregular representation of aspiration with postposed <h>

Aspirated (regular)	/k ^h p ^h t ^h tʰ/	<k ^h p ^h t ^h tʰ>	Tswana	Bantu	Botswana	(Omniglot)
Aspirated (irregular)	tʰ tʰ ^h /	tʰ tʰ ^h >				
Unaspirated (irregular)	/k ^h /	<kg>				
Unaspirated (irregular)	/f/	<sh>				

TABLE 48. Symbol choice and position of linear featurality of vowels

	Vh	Vn	Vŋ	Vx	Vq	Vɣ
Nasal vowels	✓	✓	✓			
ATR vowels	✓			✓		
Pharyngealized vowels					✓	
Back unrounded vowels						✓
Breathy vowels	✓					

TABLE 49. Symbol choice for linear featurality of click consonants

	⊖			!	‡	c	q	x
Labial clicks	✓							
Dental clicks		✓				✓	✓	
Lateral clicks			✓					✓
Aveolar clicks				✓			✓	
Palatal clicks					✓			✓

widely attested and, in some cases, are on the wane (Alaki & Norton 2015, 228).

Linear featurality is frequently applied to consonants and less often to vowels, whereas the reverse is true of multilinear featurality (Roberts in press, xx). Table 50 may give the impression of a wide range of linear strategies available for consonants other than clicks. But in fact, beyond the rather obvious strategies of spelling out secondary articulations as they sound (e.g., laterals as <Cl>, rhotics as <Cr>), and putting aside those that are infrequent and geographically limited, the choice whittles down to just two symbols: postposed <h>, and preposed or postposed apostrophe <'>. This contrasts vividly with about ten commonly used symbol-and-position choices for multilinear representations (Roberts in press, xx).

This survey has intentionally ignored one ubiquitous scenario. Linear featurality occurs whenever a primary articulation is constant across an entire series. For example, all the phonemes in the series /p^m p^w p^y p^h p' k^p/ are labial and, in a given orthography, they might plausibly be represented featurally as <p mp pw py ph p' kp>, respectively. Such

TABLE 50. Symbol choice and position of linear featurality of consonants other than clicks

	Preposed				Postposed									
	'C	mC	nC	ŋC	C'	Ch	Cr	Cw	Cu	Cy	Ci	Cl	Cv	Cgh
Prenasalization		✓	✓	✓										
Labialization								✓	✓					
Palatalization										✓	✓			
Glottalization	✓				✓	✓								
Lateralization												✓		
Dentals						✓								
Retroflexes					✓									
Aspiration						✓								
Rhoticization							✓							
Fricativization													✓	✓
Voiced plosives						✓								
Non-aspiration					✓									

cases were judged to be simply too numerous and widespread to include in the survey.

Similarly, the survey has deliberately avoided any mention of tone. Most sub-Saharan African languages are tonal, and if tone is marked at all, it is usually with a multilinear strategy involving superscript diacritics (Roberts in press, xx). Linear strategies for marking tone, on the other hand, have been proposed much more often than they have been implemented (e.g., Mann & Dalby 1987: 210; Mann 1969: 98-106; Nakuma 2002: 16; Tucker 1949: 222-223; 1964: 606). A notable exception is Côte d'Ivoire, where word-initial and -final punctuation symbols were officially endorsed for an entire generation for various Mande, Kru, and Kwa languages (ILA 1979, 20), and this is a clear case of linear featurality because the system encodes tone fully and systematically. However, more recently, this strategy has been abandoned in numerous languages in favor of multi-linear marking by means of superscript diacritics (see Roberts et al. 2019, 320 for the case of Eastern Dan). As for semiographic representations of grammatical tone (Cahill 2021: 109-111; Kutsch Lojenga 2014: 65-67; Snider 1992: 25-30), they cannot be considered featural because they encode grammar, not phonology. For these reasons, including tone in the present survey was considered unwarranted.

Finally, areal tendencies are less clear for linear than for multilinear featurality. For example, marking nasal vowels multilinearly with a subscript cedilla <Ṽ> is a distinctively Cameroonian strategy unattested elsewhere (Roberts in press, xx). However, although Cameroon has a dense concentration of languages attesting labialization, the linear strategy of representing this feature with postposed <w> is by no means restricted to that country. Granted, the linear use of postposed <x> to

mark +ATR vowels is unique to Sudan, as far as I am aware, but it is not widespread even within that country and has already been abandoned in some languages. Similarly, postposed <q> marking pharyngealized vowels linearly is a distinctively southern Africa tradition, but within that area, the languages that have this phonological feature and that use this orthographic strategy are far from numerous.

In conclusion, Table 51 summarizes the differences between linear and multilinear approaches in African Roman script orthographies.

TABLE 51. Linear and multilinear featurality in Roman script African orthographies

Linear featurality (this paper)	Multilinear featurality (Roberts in press)
• Narrow range of frequently used symbols • Frequently applied to consonants • Less frequently applied to vowels • Rarely applied to tone • Less clear areal tendencies	• Wide range of frequently used symbols • Less frequently applied to consonants • Frequently applied to vowels • Frequently applied to tone • Clearer areal tendencies

The highest priority, surely, is to empirically test the various featural strategies against one another in classroom interventions, beginning with this fundamental distinction between linear and multilinear representations. Armed with data, we will be better placed to train the next generation of orthography developers to avoid any featural strategies that do not contribute, or contribute negatively, to reading outcomes.

References

ACATBA (2005). *La phonologie et l'orthographe de la langue nzakara*. Bangui: ACATBA.

Ager, Simon (1998–2024). *Omniglot: The online encyclopedia of writing systems and languages*. Accessed 11 March 2024. URL: <http://www.omniglot.com/>.

Alaki, Thomas Kuku and Russell Norton (2015). “Kadaru-Kurtala Phonemes”. In: *Dotawo: A Journal of Nubian Studies* 2.10, pp. 215–230.

Allison, Sean (2006). *Alphabet et orthographe de Kokoko de Makary (Mpadî): Version préliminaire*. Yaoundé: SIL Cameroun et Ministère de la Recherche Scientifique et de l’Innovation.

Alomofor, Christian and Stephen C. Anderson (2005). *Awing orthography guide*. Yaoundé: SIL Cameroon.

Anderson, Stephen C. (2014). *Isu orthography guide*. Yaoundé: SIL Cameroon, Ministry of Scientific Research, and Innovation.

Anonymous (2013). *Lugbarati orthography guide*. Institution unknown.

- (2015). *A guide prepared at the Gworog orthography workshop held at the Gworog literacy and translation project office, Kagoro*. Kagoro, Kaduna State: Gworog language Kaura LGA.
- Armstrong, R. G. (n.d.). *Idoma orthography*. Institution unknown.
- Bamata, Subama (1997). *Manuel d'orthographe Budu*. SIL Democratic Republic of Congo.
- Barnwell, Katharine (n.d.). *The orthography of Mbembe, a Cross River language*.
- Bezrukov, Nikita and Ronan Soni (2023). "Featurality as a byproduct of script inheritance". Paper presented at the 14th workshop of the Association of Written Language and Literacy, Temple University, Rome.
- Bilal, Clement Murba Wau (2004). "The Adamawa-Ubangi languages in Sudan". In: *Occasional Papers in the Study of Sudanese Languages*. Ed. by Leoma Gilley. Entebbe: SIL Sudan, pp. 175–184.
- Boyd, Ginger (2015a). *Précis d'orthographe de langue nulibie*. Yaoundé: Ministère de la Recherche Scientifique et de l'Innovation & SIL Cameroun.
- (2015b). *Précis d'orthographe de langue numála*. Yaoundé: Ministère de la Recherche Scientifique et de l'Innovation & SIL Cameroun.
- (2016). *Précis d'orthographe de langue nuasue (yangben)*. Yaoundé: SIL Cameroun.
- Broe, Michael (1992). "An introduction to feature geometry". In: *Papers in Laboratory Phonology*. Ed. by G. J. Docherty and D. R. Ladd. Cambridge: Cambridge University Press.
- Cahill, Michael (2021). "Marking grammatical tone in orthographies: Issues and challenges". In: *Celebrating 50 years of ACAL: Selected papers from the 50th Annual Conference on African Linguistics*. Ed. by Akinbiyi Akinlabi et al. Berlin: Language Science Press, pp. 23–34.
- CISOLA (2010). *Guide de l'orthographe de l'ajagbe*. Azové: Comité International du Suivi de l'Orthographe de la Langue Aja.
- Clements, George N. (1985). "The geometry of phonological features". In: *Phonology Yearbook*, pp. 225–252.
- Cox, Bruce (2005). *Kemezung orthography guide*. Yaoundé: SIL Cameroon.
- Davison, Phil (2009). *Web orthography guide*. Yaoundé: Ministry of Scientific Research, Innovation, and SIL Cameroon.
- Diatta, Lena (2013). *Guide d'orthographe de la langue gusilay: version finale. Guide pratique pour l'apprentissage et la découverte de la langue gusilay par écrit*. Dakar et Thionck-Essyl: SIL Sénégal et Coordination Locale des Opérations en Alphabétisation (CLOA).
- Drame, Bandiougou S. and Marianne Hagg (2020). *Guide d'écriture de la langue soninké: Sooninkanxannen safandimoxon geesundidokko SIL Sooninkanxannen Gollinjananon Sapa*. Bamako: SIL Mali.
- Duerksen, John (2004a). "A journey in orthography evaluation: My Otuho experience". In: *Occasional Papers in the Study of Sudanese Languages*. Ed. by Leoma Gilley. Entebbe: SIL Sudan, pp. 39–56.

- Duerksen, John (2004b). "Dinka-Nuer orthographies". In: *Occasional Papers in the Study of Sudanese Languages*. Ed. by Leoma Gilley. Entebbe: SIL Sudan, pp. 69–76.
- Eyoh, Julius A. and Robert Hedinger (2008). *Mfumte orthography guide*. Yaoundé: National Association of Cameroonian Language Committees (NACALCO).
- Eyoh, Julius A. and Echebi Emmanuel Sandamu (2009). *Mbembe orthography guide*. Yaoundé.
- Fadul, Yousif Aljuzuli et al. (2020). *Jumjum—English Dictionary*. Dallas, TX: SIL International.
- Fallon, Paul D. (2006). "Blin orthography: A history and an assessment". In: *Selected proceedings of the 36th Annual Conference on African Linguistics*. Ed. by Olaoba F. Arasanyin and Michael A. Pemberton. Somerville, MA: Cascadilla Proceedings Project, pp. 93–98.
- Gilley, Leoma (2004). "The Lwoian family". In: *Occasional Papers in the Study of Sudanese Languages*. Ed. by Leoma Gilley. Entebbe: SIL Sudan, pp. 165–174.
- Gnanadesikan, Amalia (2023). "The representation of distinctive features in writing systems: Omission and commission". Paper presented at the 14th workshop of the Association of Written Language and Literacy, Temple University, Rome.
- Grove, Dan and Jennifer Wright (2012). *Bambalang (Chrambo) orthography guide*. Yaoundé: SIL Cameroon, Ministry of Scientific Research, and Innovation.
- Hall, Ed and Marion Hall (2004). "Kudugli-Krongo". In: *Occasional Papers in the Study of Sudanese Languages*. Ed. by Leoma Gilley. Entebbe: SIL Sudan, pp. 57–68.
- Hamm, Valerie and Cameron Hamm (2016). *Bafanji (Chufie') orthography guide*. Yaoundé: Ministry of Scientific Research, Innovation, and SIL Cameroon.
- Hartell, Rhonda L. (1993). *Alphabets of Africa*. Dakar: UNESCO & SIL-Senegal.
- Hedinger, Robert (2011). *Akoose orthography guide*. Yaoundé: SIL Cameroon et al.
- Hellenthal, Anne-Christie and Constance Kutsch Lojenga (2011). *Phonology / Orthography statement for the Gwama language*. Dallas, TX: SIL International.
- Hubert, Jacques and Paul Achille Mavoungou, eds. (2010). *Écriture et standardisation des langues gabonaises*. Stellenbosch: Sun Press.
- ILA (1979). *Une orthographe pratique des langues ivoiriennes*. Abidjan: Institut de Linguistique Appliquée, Université d'Abidjan.
- Iornumbe, Charles (2020). "The development of a standard and practical orthography for Tiv language". In: *Journal of the Nigerian Languages Project 2*, pp. 47–58.

- Joswig, Andreas et al. (2010). *Statement of working orthography [for Borna]*. Ed. by Benishangul-Gumuz Language Development Project. Assosa, Ethiopia.
- Kilpatrick, Eileen (2004). "Orthographies of Moru-Ma'di languages". In: *Occasional Papers in the Study of Sudanese Languages*. Ed. by Leoma Gilley. Entebbe: SIL Sudan, pp. 85–91.
- Kinnaird, William (2005). *Proposition d'orthographe pour la langue vamé*. Yaoundé: Ministère de la recherche scientifique et de l'innovation et SIL Cameroun.
- Kutsch Lojenga, Constance (2014). "Orthography and tone: A tone system typology with implications for orthography development". In: *Developing orthographies for unwritten languages*. Ed. by Michael Cahill and Keren Rice. Dallas: SIL International, pp. 49–72.
- Langhout, Bep (2022). *Sari orthography guide*. Yaoundé: SIL Cameroon.
- Legbeti, Agnes Temitope (2016). *An orthography of the Ọsòsò language*. Benin City, Nigeria: University of Benin.
- Linton, Pauline (2016). *L'orthographe du latege: Edition préliminaire*. Léconi, Gabon: Global Outreach Mission.
- Lux, Cindy (2018). *Nooni orthography guide*. Yaoundé: SIL and Nooni Language Committee.
- Madugu, I. S. G. (n.d.). *Nupe orthography*. Institution unknown.
- Mann, Michael and David Dalby (1987). "Writing African languages". In: *A thesaurus of African languages: A classified and annotated inventory of the spoken languages of Africa with an appendix on their written representation*. London: Hans Zell Publishers, pp. 206–219.
- Mann, W. M. (1969). "Tone-marking an African language, with application to Bemba". In: *African Language Review* 8, pp. 98–107.
- McLean, Greg L. (2020a). *Koffa orthography guide*. Yaoundé: Cameroon Association for Bible Translation and Literacy (CABTAL).
- (2020b). *Kwaja orthography guide*. Yaoundé: Cameroon Association for Bible Translation and Literacy (CABTAL).
- (2020c). *Lidzonka orthography guide*. Yaoundé: Cameroon Association for Bible Translation and Literacy (CABTAL).
- (2020d). *Ndə Wuli orthography guide*. 3rd ed. Yaoundé: Cameroon Association for Bible Translation and Literacy (CABTAL).
- Mfouhou, Chrisnah Renaudot (2022). *Petit guide d'orthographe et de grammaire luumbu parlé dans le Niari en République du Congo*. Brazzaville: SIL-Congo.
- Mkounga, Tala Blaise (2022). *Mungong orthography guide*. Yaoundé: SIL Cameroon.
- (2023). *Naami orthography guide*. Yaoundé: SIL Cameroon.
- Murrell, Paul (2020). *Précis d'orthographe du ngando*. Bangui: ACATBA et SIL International.
- Nakuma, Constancio K. (2002). "Guide to the Dagaarhre Orthography". In: *Journal of Dagaare Studies* 2, pp. 1–43.

- Ngulube, Isaac Eyi (2011a). *Ekpeye orthography*. Abuja: Nigerian Educational Research & Development Council (NERDC).
- (2011b). *Ogba orthography*. Abuja: Nigerian Educational Research & Development Council (NERDC).
- Njeck, Mathaus M. and Stephen C. Anderson (2010). *Bakoko orthography guide*. Yaoundé: Cameroon Association for Bible Translation and Literacy (CABTAL).
- Njeck, Mathaus Mbah (2008). *Moghamo Orthography Guide*. Yaoundé, Cameroon: SIL.
- Njeck, Mathaus Mbah and Robert Hedinger (2006). *Baba 1 orthography guide*. Yaoundé: SIL Cameroon.
- Nzogi, Richard (2011). *Empandiika ya Ruruuli-Runyala: Ruruuli-Runyala orthography/spelling guide*. Nakasongala, Uganda: Ruruuli-Runyala Language Society, Luke Partnership Project and SIL Uganda.
- Odom, Shannon Ronit (2015). *Zinza orthography statement*. Dar Es Salaam: SIL Uganda-Tanzania.
- PanSALB (2019). *Milayo ya kupelele na kunwalele kwa Tshivenda*. Arcadia: Pan South African Language Board.
- Persson, Janet (2004). “Bongo-Bagirmi languages in Sudan”. In: *Occasional Papers in the Study of Sudanese Languages*. Ed. by Leoma Gilley. Entebbe: SIL Sudan, pp. 77–84.
- Prinz, Angela (2011). *Orthography for the Massalit language*. Ndjamen, Cameroon: SIL Chad.
- Pweddson, Nicholas Nzovudeto (n.d.). *Bwatye orthography*. n.d.
- Roberts, David (2008). “The two Kabiye orthographies: a sociolinguistic and linguistic comparison”. In: *Written Language & Literacy* 11.1, pp. 49–72.
- (in press). “Multilinear featurality in Roman script African orthographies”. In: *Handbook of Non-linear writing systems*. Ed. by Heather Winskel. Cham: Springer.
- Roberts, David, Dana Basnight-Brown, and Valentin Vydrin (2019). “Marking Tone with Punctuation: Orthography Experimentation and Reform in Eastern Dan (Côte d’Ivoire)”. In: *Proceedings of Graphemics in the 21st Century, Brest 2018*. Ed. by Yannis Haralambous. Brest: Fluxus Editions, pp. 293–327.
- Sampson, Geoffrey (1985). *Writing systems: A linguistic introduction*. Stanford, CA: Stanford University Press.
- Sande, Hannah and Madeleine Oakley (2023). “A typological survey of the phonological behavior of implosives: Implications for feature theories”. In: *Phonological Data and Analysis* 5.7, pp. 1–46.
- Scherrer, Elaine Marie (2014). *L’orthographe guemzek : version révisée*. Yaoundé: SIL Cameroon.
- Schroeder, Leila (2008). *Bantu Orthography Manual*. Dallas: SIL International.

- SIL (2001). *Proposition d'une orthographe de la langue toussian*. Ed. by SIL Burkina Faso. Ouagadougou.
- (2011). *Karon orthography: An explanation and justification of the rules of orthography in the Karon language*. Dakar: SIL Senegal.
- (2012). *Phonology & orthography sketches of South Sudanese languages*. Juba.
- (2017). *Apprenons à lire notre langue maintenant*. Ouagadougou.
- Smith, Tony (2001). *Alphabet et orthographe muyang*. Yaoundé: SIL Cameroon.
- Snider, Keith L. (1992). "Grammatical tone and orthography". In: *Notes on Literacy* 18.4, pp. 25–30.
- Stirtz, Timothy M. (2014). *Bēli phonology, tone and orthography*. Dallas, TX: SIL International.
- Tabb, Waller C. et al. (2011). *Kitabo kya kubabula kubandiika Lubwisi: The Lubwisi spelling guide*. Bundibugyo: The Kwamba/Lubwisi Development Association and SIL International.
- Tench, Paul (n.d.). *A new orthography in Tera*. Cardiff: Cardiff University Centre for Language and Communication Research.
- Tucker, A. N. (1949). "Sotho-Nguni orthography and tone-marking". In: *Bulletin of the School of Oriental and African Studies, University of London* 13.1, pp. 200–224.
- (1964). "Systems of tone-marking African languages". In: *Bulletin of the School of Oriental and African Studies, University of London* 27.3, pp. 594–611.
- Tuinstra, Anna (2018). *Revised Ipulo orthography guide*. Yaoundé: SIL Cameroon, Ministry of Scientific Research, and Innovation.
- Viljoen, Mélanie et al. (2009). *Précis d'orthographe pour la langue buwal: édition préliminaire*. Yaoundé: SIL Cameroun.
- Vydrin, Valentin, Emmanuel Zeh, and Gué Nestor (2019). *Guide de nouvelle orthographe dan de l'est pour les apprenants familiers avec l'ancienne orthographe*. Man: EDILIS.
- Weber, Joan and Martin Weber (2008). *Précise d'orthographe pour la langue kwanja (dialecte Sùndàn)*. Yaoundé: Église Évangélique Luthérienne du Cameroun et Association Nationale des Comités de Langues Camerounaises.
- Wiese, Richard (2011). "The representation of rhotics". In: *The Blackwell companion to phonology*. Ed. by Marc van Oostendorp et al. Oxford: John Wiley, pp. 1–19.
- Williamson, Kay (1984). *Practical orthography in Nigeria*. Ibadan: Heinemann Educational Books.
- Yigezu, Moges and Binyam Sisay Mendsu (2015). "The orthography of Hamar". In: *Studies in Ethiopian Languages* 4, pp. 1–16.
- Yip, May (2004). "Phonology of the These language". In: *Occasional Papers in the Study of Sudanese Languages*. Ed. by Leoma Gilley. Entebbe: SIL Sudan, pp. 93–117.

The Use of Brackets in German (And Beyond)

Jonas Romstadt & Lorenz Vielhauer

Abstract. Brackets are established punctuation units in German. Nevertheless, little is known about the conditions under which writers employ this sign. This serves as the starting point for this study. Based on an analysis of over 1,000 Abitur examination essays (the final school-leaving exam in Germany), we demonstrate that usage-based prototypes can be distinctly characterized. These pertain, first, to the form of the sign, second, to its function, and third, to the linguistic structure(s) in which brackets are employed. In a subsequent step, we contrast the resulting profile of the brackets with the plus sign in live tickers—a unit that has thus far been rarely (if ever) described in the context of punctuation. Finally, these usage-based insights are contextualized within a grapho-linguistic framework.

1. Introduction: What Is This All About?

The novel *Berlin Alexanderplatz* by Alfred Döblin was published in 1929. Over the course of the plot, the reader accompanies Franz Biberkopf, a wage laborer who is released from prison and is now trying to build a new life for himself. In 2002, it was voted to be one of the 100 best books of all time.¹ The novel begins with the story of Franz Biberkopf taking the tram from the prison to the city of Berlin:

“Er stand vor dem Tor des Tegeler Gefängnisses und war frei. [...] Der Aufseher am Tor spazierte einige Male an ihm vorbei, zeigte ihm seine Bahn, er ging nicht. Der schreckliche Augenblick war gekommen (schrecklich, Franze, warum schrecklich?), die vier Jahre waren um. Die schwarzen

Jonas Romstadt

University of Oldenburg, Ammerländer Heerstraße 114-118, 26129 Oldenburg, Germany

E-mail: jonas.konstantin.romstadt@uni-oldenburg.de

Lorenz Vielhauer

University of Bonn, Breite Straße 97, 53111 Bonn, Germany

E-mail: lorenzvielhauer@web.de

1. Cf. <https://www.theguardian.com/world/2002/may/08/books.booksnews> (accessed: 19.11.2024).

eisernen Torflügel, die er seit einem Jahre mit wachsendem Widerwillen betrachtet hatte (Widerwillen, warum Widerwillen?), waren hinter ihm geschlossen." (Döblin 1947 [1929]: 12)

"He stood in front of the gate of Tegel Penitentiary and was a free man. [...] The guard at the gate walked past him a few times, showed him his way to the tram, he didn't go. The terrible moment had come (terrible, why terrible, Franz?), the four years in prison were over. The black iron gates, which he had been looking at for a year with growing disgust (disgust, why disgust?), were closed behind him."

The third-person narrator reports on the discomfort Franz Biberkopf feels as he stands in front of the prison and should actually be starting the next phase of his life on a positive way. Analyzing this inner conflict was the subject of German school leaving exams (*Abiturprüfung*) in 2008. The students were given the following task:

Interpret this text excerpt from "Berlin Alexanderplatz" by Alfred Döblin.

A number of linguistic peculiarities can be analyzed and interpreted in the given text. Some students focused on the two insertions in brackets that appear in the text (<(terrible, why terrible, Franz?)> and <(disgust, why disgust?)>), see for example (1).

- (1) Zusätzlich nutzt der Autor durch [sic!] Zugabe in Klammern [...], um seine Auffassung deutlich zu machen. (German exam 2008, grade C)
In addition, the author uses additions in brackets [...] to insert comments in order to make his view clear.

This statement by a female student has two sides: On the one hand, it discusses a formal aspect: the brackets are *inserted into* the surrounding sentence and not placed at its edge. On the other hand, the author interprets them functionally: the author of the text *reveals himself* in his role and makes a subjective comment—similar to a comment from the off (*to make his view clear*).

This assessment is not so far away from what linguists have said about brackets. Geoffrey Nunberg, for example, states:

Most often, however, the features of the context of production that are relevant to interpretation will involve the author's intentions in inscribing the text, and will be construed as providing a kind of metalinguistic comment on the form or content of the associated antecedent. (Nunberg 1990: 115–116)

In (1), the author has also identified this metalinguistic comment. It is not free, but refers to the context. Behrens (1989: 104) assigns bracketed units a loose "contextual reference"² to the surrounding construction. This is also the case in the given extract from *Berlin Alexanderplatz*: the

2. All German quotes in this paper were translated by the authors.

comments are inserted into the surrounding syntactic structure. However, they are irrelevant for the overall understanding. If the texts inserted between the brackets were omitted, the sentence would be complete both syntactically and semantically. Gallmann (1985: 167) therefore analyses texts in brackets as “parts of the text with the characteristics ‘incidental’, ‘dispensable for the overall understanding’, and therefore ‘can be omitted’”. Zimmermann (1968: 25) similarly states that the usage of brackets is a matter of “marking information as incidental”.

What all these commentaries on brackets have in common is that they focus on the *systematic* function of the sign. They discuss when brackets can *possibly* be used more than when they are *actually*.

What, therefore, only plays a subordinate role is the question of when exactly brackets are written—and, related to this, which functional (and formal) potentials of the brackets are used to what extent. If the usage is explicitly addressed at all in the mentioned studies, they often only work with small, apparently intuitively compiled collections of evidence. An analysis of the status quo of the usage of the brackets is, as well as for punctuation in general, a research desideratum (cf. Fürstenberg 2023: 4; Langlotz 2017: 151).

This is the starting point of the present study. A large text corpus of school leaving exams will be used to determine *when* writers actually use brackets, *how* they do so, and *why*.

In detail, the text is structured as follows. First, the brackets are described as a punctuation unit³ and classified in the punctuation system of German. Here, a distinction is made between the reader’s and the writer’s perspective (Section 2). Building on these findings, the actual use of the brackets in school leaving exams is described. We will give reasons why we believe that these texts are actually a rather poor data basis for this endeavor—but nevertheless the best option available (Section 3). We will also trace developments over time that relate to the form of the brackets, their function, and the variation that can be observed (Section 4). We will show that the brackets have gradually developed their synchronous functional specificity in the school leaving exams.

Building on these findings, the view is broadened in Section 5 and the formal-functional profile of brackets in German is compared to the plus sign <+>. We will show that it has a mirror-image function in certain, well-defined communicative situations. The article concludes with an outlook (Section 6). Overall, it is a plea for usage-based grapholinguistics and comparative analyses of writing systems, which have so far been the exception in the field of punctuation.

3. We use the term punctuation unit rather than punctuation sign or punctuation mark in order to adopt a neutral term that equally denotes both the form and the function of the respective unit (cf. Meletis & Romstadt, accepted).

2. Brackets as Part of the (German) Punctuation System

Punctuation is a written phenomenon realized through segmental means. Building on the analysis by Bredel (2008; 2020), it can be further delineated as follows: Punctuation units differ from other segmental elements of a writing system in that, on the one hand, they are identifiable even without a graphic context, yet, on the other hand, they are not associated with conventionalized vocal gestures during reading (i.e., they are not recodable).

The first aspect distinguishes them from the space, which, therefore, is not considered a punctuation sign (contrary to, e.g., Jones 1995). The latter aspect differentiates punctuation units from other segmental means of a writing system.

This holds true regardless of the specific writing system under consideration. Specific to German, the criterion of recodability differentiates punctuation units from letters, numerals, and special characters. The latter are commonly associated with vocal gestures, such as <%> being realized as [pʁo'ʦɛnt] (“percent”) (issues of delineation are discussed in Chapter 5, as well as in Bredel 2008).

Based on this definition, the following 12 units can be identified as punctuation units in German: <., : ; „“ () ! ? — ‘ ’ ... >.⁴

Two of these hold a special status: quotation marks and brackets. They are paired signs, which constitute a necessary condition: They *always* consist of two elements. This differs from cases where, for instance, commas or dashes are used in pairs. Using only one bracket or only one quotation mark constitutes an orthographic error (see Section 4.2 for exceptions). Their pairing is obligatory, and they always consist of an initial and a final element that together constitute a single punctuation unit.

Structurally, this can be explained by the organization of written language. Bredel (2008: 92) defines this in terms of “prototypical written-grammatical constituent structures of punctuation phrases.” These are descriptions of written word phrases (WWP) in their interaction with other graphically realized segmental means of writing. For brackets, this results in Figure 1.

Brackets thus encompass a written word sequence within their scope. This is also true in the above-cited excerpt from *Berlin Alexanderplatz*. It should be noted that, theoretically, even a single graphematic word can

4. Note that these punctuation units are abstracted from their actual form. They are more or less functional labels for concrete forms. Some of these punctuation units may appear in different forms, such as the brackets (<()> vs. <[]> vs. << >> vs. <>, see section 4.1). This form of variation is excluded here while creating the inventory, as all of these (allographic) forms can be assigned to the same punctuation unit in the given inventory.

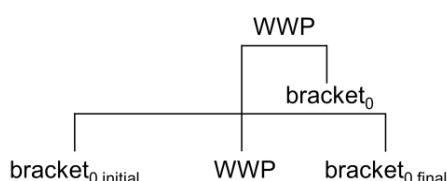


FIG. 1. Constituent structure of written word phrases (WWP) with brackets (cf. Bredele 2008: 92)

be enclosed in brackets (2). Similarly, it is possible to place only parts of a graphematic word (3) or other punctuation units (4) in them.

- (2) Sie hatten feste Zielvorstellungen (Demokratie) ... (German exam 1978, grade C)
They had clear objectives (democracy) ...
- (3) ... diese gescheiterte(n) Beziehung(en) ... (German exam 2018, grade A)
... these failed relationship(s) ...
- (4) ... da ein Mensch mit seiner geistigen Entwicklung nie (!) fertig werden kann (German exam 1958, grade C)
... since a person can never (!) fully complete their intellectual development

Brackets demonstrate considerable flexibility in this regard, functioning as a kind of “all-inclusive-signs.”

In Figure 1, however, a second aspect is also relevant: the function of brackets emerges only in light of the interplay between punctuation units and their surrounding co-text. This observation itself has two dimensions: on the one hand, it concerns the linguistic structure within the brackets; on the other, it pertains to the carrier structure outside them.

In the structure described in Figure 1, brackets have long been part of the German writing system. Masalon (2014) identifies early instances in texts written between 1470 and 1510 (19 bracket pairs in 9 texts; cf. Masalon 2014: 247–250). Parkes (1993: 48) traces brackets even further back to the late 14th century. In early examples, brackets are predominantly associated with a circular form: they are “presumably (at least conceptually) remnants of a circle, which was used to enclose specific words and thus set them apart from the surrounding text” (Klein 1998: 183). This explains why brackets are still used in pairs today and why they continue to visually evoke a circle, at least in their rounded form, which is also the oldest.

The introduction of brackets into the German writing system led to their early high frequency (cf. Höchli 1981: 19). They benefited, on the one hand, from the overall increasing syntactic complexity in German at that time, which was enabled by the growing pervasiveness of written language in society. On the other hand, they were also a product of these

developments, as they facilitated the introduction of more complex syntax (cf. Klein 1998: 183). Brackets, thus, are a symptom as well as a driver of these processes.

This dual role was recognized and discussed metalinguistically at an early stage. In the first metalinguistic analyses of German punctuation, brackets were often discussed in detail—frequently more so than any other punctuation mark (a status that today would likely belong to the comma) (cf. Höchli 1981: 290). These spacious discussions were often accompanied by stylistic and linguistic criticisms warning against excessive use of brackets. For example, Gottsched (1766: 38) remarked: “Since it is not aesthetically pleasing to make too many insertions, one should prefer to use simple dashes for shorter phrases.”⁵

This is no longer the case today. In the current official rules of the German Orthographic Council (*Rat für deutsche Rechtschreibung*) (2024), brackets are addressed only superficially in a single paragraph: “Brackets indicate the presence of supplementary information” (§ 76).

Two aspects are noteworthy here: First, this statement is formulated from the perspective of readers. It does not address what writers do when (or why) they use brackets but rather discusses the effect this sign has on readers (cf. Bredel/Wöllstein 2024). Second, only a single primary function of brackets is identified. There is no acknowledgment of poly-functionality, which has been proposed by Zifonun et al. (1997: 296).

The explanatory notes to § 76 of the official orthographic rulebook further state: “Brackets contain clarifications, additions, or explanations. The bracketed addition is always omissible. The remaining structure is grammatically and semantically complete.”

This elaborates on what was previously described as the interplay between the text within the brackets and the surrounding structure (see Figure 1). By assigning the property of ‘omissibility’ to the bracketed unit, brackets are framed as a tool of linguistic economy. At the same time, this explains why brackets primarily assume a structural function in texts (and words): “Only those units that do not contribute to the textual representation of a factual proposition can be bracketed” (Bredel 2008: 142).

These units may either be integrable into the surrounding syntactic structure (5) or not (6).

- (5) An Kleidung ist (fast) gar nicht zu denken. (German exam 1949, grade D)
Clothing is (almost) out of question.
- (6) Er (der Leser) muß nämlich einen verschachtelten Satz solange im Gedächtnis behalten, bis der Verstand zum Schluß das ganze „Rätsel“ auflöst. (German exam 1963, grade D)

5. The specific form of brackets is of secondary importance to Gottsched. He notes that round brackets are the most common but adds: “Others make them like this [=]. Both are the same” (Gottsched 1766: 38).

He (the reader) must keep a convoluted sentence in memory until the mind finally resolves the entire "riddle" at the end.

In the case of (5), Bredel (2020: 61) refers to an integrable *construction bracket*. She terms (6) a *commenting bracket*. If the brackets in the latter are omitted, this would result in an orthographic error:

- (5') An Kleidung ist fast gar nicht zu denken.
Clothing is almost out of question.
- (6') *Er der Leser muß nämlich einen verschachtelten Satz solange im Gedächtnis behalten, bis der Verstand zum Schluß das ganze „Rätsel“ auflöst.
**He the reader must keep a convoluted sentence in memory until the mind finally resolves the entire "riddle" at the end.*

Both (5) and (6) share the characteristic that the truth-value of the surrounding structure is not influenced by the elements in brackets. It remains unchanged even if the parts inside the brackets and the punctuation unit itself are omitted.

This also explains why brackets confer a special status upon the text within them: "Brackets elevate the information inside [...] to a special information status *with respect to the information of the carrier structure*" (Baumgarten, Meyer & Özçetin 2008: 187, our emphasis).

The final restriction is crucial: Central to the analysis is the relationship between the elements inside and outside the brackets. This is why brackets can be used to establish a sub-discourse within a text, allowing authors to express themselves in their role. The writer shifts from being a "covert" to an "overt" writer (cf. Bredel 2008: 149), revealing themselves—within the brackets:

[This is] a method of stressing the presence of a writer in the text who is involved in information organization, as brackets signal that different pieces of information in the sentence are given different information statuses. (Baumgarten, Meyer & Özçetin 2008: 188)

This may also explain why brackets are frequently used in academic texts (cf. Biber & Gray 2016: 120). They graphically license the introduction of a potentially subjective sub-discourse.

Up to this point, the focus has been on characterizing the systematic functions of brackets (in German).

A related but distinct issue concerns actual usage: when and where are brackets employed, and to what extent do competent writers utilize their functional nuances? Initial insights into this can be found in two corpus studies that primarily focused on translation practices.

Baumgarten, Meyer & Özçetin (2008) examine German popular science texts and their English translations. In the German original texts, they identify 284 pairs of brackets, which they classify into 13 categories. Brackets are most commonly used for "reader-oriented exophoric and intertextual reference" (107 tokens, 37.7%), followed by 36 instances

each of “specialized terminology” and “geographical information” (12.7% each) (cf. Baumgarten, Meyer & Özçetin 2008: 192). This highlights what was previously noted about brackets: they create a sub-discourse to explicate supplementary information. At the same time, and closely related, brackets in (popular) science texts provide a way for authors to directly address readers, more or less explicitly, in their role as writers.

More recent data can be found in the larger corpus study by Levin & Ström Herold (2021), who analyzed 5,923 pairs of brackets in German, Swedish, and English texts and their respective translations. Of particular interest for the present study are the 539 pairs of brackets in the eight German original texts.⁶

Levin & Ström Herold functionally distinguish between *content-oriented* and *interpersonal* brackets:

As the terms imply, content-oriented brackets focus on adding to the content of the text, while interpersonal ones are more reader- or author-oriented and as such may be considered a prototypical case of using brackets as communicative marks. (Levin & Ström Herold 2021: 127)

Each class is further subdivided. Among content-oriented brackets, there are those used to indicate *synonyms* for a “co-referential [...] item outside the brackets” (Levin & Ström Herold 2021: 127), as well as the broader category of *specification*, which provides omissible additional information about the carrier structure.

By contrast, interpersonal brackets focus more explicitly on the relationship between author and reader. This is most evident in the category of *subjective author comment*, where writers subjectively evaluate or assess parts of the carrier structure (or the structure as a whole). The interpersonal character of brackets is also clear in the subcategory of *reader address*, which involves “instances where the reader is addressed ‘outside the text’, [...] and metatextual comments guiding the readers” (Levin & Ström Herold 2021: 128). Finally, there is the interpersonal subcategory of *hedge*, described by Levin & Ström Herold (2021: 128) as “in some ways categorically ambiguous.” This ambiguity primarily concerns its distinction from the content-oriented category of *specification*. However, when it involves subjective qualifications where the author is directly recognizable as the qualifying agent, an interpersonal classification seems justified.

In the data analysis in Section 4, a conservative annotation is applied: The category hedge is only assigned where an interpersonal character is

6. The relevant findings of Levin & Ström Herold (2021) on the translation practices of brackets are not further elaborated here. Overall, they state: “Most brackets are retained in translations, and there are systematic target-language-specific preferences for the retention rates” (Levin & Ström Herold 2021: 134).

explicitly expressed linguistically. The overall distribution in Levin & Ström Herold's (2021) corpus is as follows:

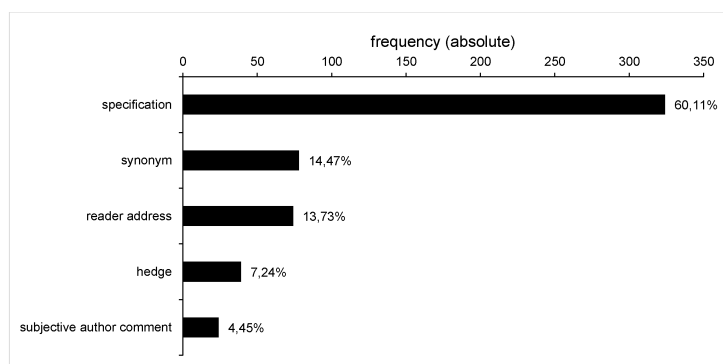


FIG. 2. Bracket functions in German texts (cf. Levin & Ström Herold 2021: 129)

Brackets are primarily used to add specifying supplementary information to the surrounding carrier structure, in a content-oriented manner. By contrast, interpersonal brackets make up only a smaller portion. The function reader address, which most closely corresponds to the most prominent subcategory identified by Baumgarten, Meyer & Özçetin (2008), is still the most frequent subcategory here. This is an initial indication of a genre-specific use of brackets.

3. Examining Punctuation Use

The basic mode of writing itself is variation. This is, grapholinguistically, a “largely understudied topic” (Meletis 2020a: 105), what is undoubtedly (and even more so) the case for punctuation usage (cf. Rössler, Besl & Saller 2022).

When specific punctuation usage has been linguistically described, it has often been done in large newspaper corpora. Studies of this kind can provide relevant insights due to the data volume, but press texts potentially undergo an editorial process, and the influence of AI text generation is not transparent. Thus, such corpus studies essentially reflect the “combined” usage of writers and editors.

Unedited texts are included in large web corpora, which refer to, for example, blogs or digitally mediated communication (cf. Busch 2021). These corpora, however, due to technical development, have only a little or no temporal span. It is simply impossible to draw conclusions about bracket usage in the early 20th century using web corpora.

Learner corpora seem to be better suited, but only show milestones in the acquisition process of writers.

The best compromise among all of these limitations seems to be school leaving exams. They have been written in Germany under comparable conditions for over 100 years. Authors have a set amount of time to handwrite a text without any additional tools (other than a dictionary). In the “world of written texts,” they emerge in fully-regulated orthographic spaces (cf. Sebba 2007: 43). This is directly related to the high social obligation perceived in German orthography. Writers of school leaving exams are anxious to produce orthographically correct texts because spelling mistakes are officially sanctioned. They therefore best demonstrate what can be understood as the authentic writing skills of more or less ‘competent’ writers at the end of their school years (meaning that most of them will no longer attend institutionalized spelling classes) (cf., e.g., Ransmayr 2022, Eiben-Zach 2023).

3.1. The GraphVar-Corpus

The use of brackets is to be examined in the GraphVar corpus for these reasons. This corpus contains over 1,800 school leaving exams written at a secondary school in Lower Saxony (cf. Romstadt, Strombach & Berg 2024).

Here, a portion of this corpus is described, specifically, only those exams that were written in the subject of German. This is the only subject that is present in all corpus years, and historically, the German essay holds a special position (cf. Reh, Kämper-van den Boogaart & Scholz 2017). These are 1,027 exams with 1,363,029 tokens. All units of the school leaving exams that are quoted and thus attributed to other authors by the writers are excluded (cf. Berg, Romstadt & Neitzert, 2021). The exams are not evenly distributed across the individual years, see Figure 3.

3.2. Methodological Remarks

In addition to the imbalance in terms of the number of texts described in Figure 3, there are other factors that limit the comparability of the texts. For instance, the exams have generally become longer over time. An average exam in the year 1917 was 713 tokens long, whereas 101 years later, in 2018, the average length was 2,087 tokens. To compensate for this, relative values are mainly provided below. These results are less sensitive to fluctuations in text length.

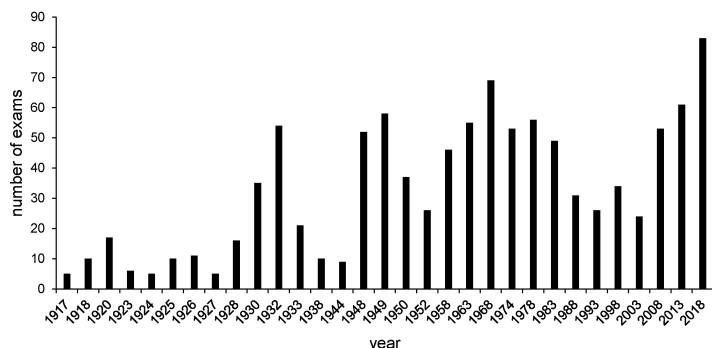


FIG. 3. Number of exams in the corpus, aggregated by year of origin

The historical comparability is also limited by two factors. These are (a) changes in the task environment and (b) sociodemographic developments.

Ad (a): In the early years, the explicit goal of final exams was not to assess subject-specific competencies, but to measure general academic ability (earlier, the term “Reifeprüfung” ‘maturity examination’ rather than “Abiturklausur” ‘graduation exam’ was used). This also means that in the early years, students were often expected to address subjective value issues. A comparative example of two task formulations makes this immediately clear (see Table 1). In 1948, a so-called reflection essay (*Besinnungsaufsatz*) was expected (cf. Ludwig & Merchert 1987), which was thought to reveal the eloquence and argumentation skills of the writers. Today, a more or less scientific treatise on a subject-related topic is required.

Ad (b): In the past, taking the school leaving exam was the privilege of a small portion of the (male) population. Today, more than 50% of the German population takes the corresponding exam. Sociologically, this has been described as “educational expansion” (cf. Bölling 2010).

Both aspects need to be considered in the further analysis of the data presented here, and we will return to them in the analysis section.

Our quantitative study is methodologically relatively simple. In the GraphVar corpus, we query all instances of brackets that do not appear in quotes. In total, this amounts to 17,572 brackets (initial and final brackets are counted separately here, cf. Section 4). These tokens will be further analyzed and linguistically described.

TABLE 1. Exemplary task formulations

1948	2008
Welche Wünsche und Gedanken kommen mir beim Anblick eines Großflugzeuges? <i>What wishes and thoughts come to my mind when I see a large airplane?</i>	1. Interpretieren Sie den vorliegenden Textauszug aus „Berlin Alexanderplatz“ von Alfred Döblin. 2. Setzen Sie sich mit der Frage auseinander, inwiefern die Figurenkonzeptionen von Franz Biberkopf und Lenz unter dem Aspekt von Ich- und Wirklichkeitszerfall miteinander vergleichbar sind. <i>1. Interpret the following excerpt from “Berlin Alexanderplatz” by Alfred Döblin. 2. Discuss to what extent the character concepts of Franz Biberkopf and Lenz can be compared under the aspect of decomposition of reality and ego.</i>

4. The use of brackets in German

First of all, the frequency of the tokens <(> and <)> is evaluated in the GraphVar corpus. They are counted separately, although, as mentioned above, they form a single punctuation unit. The absolute value is normalized to the total corpus size, so the frequencies are determined per 1,000 tokens here.

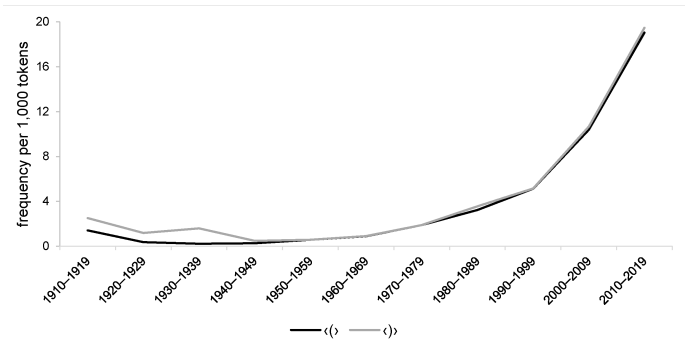


FIG. 4. Frequency of round brackets per 1,000 Tokens over time

It is particularly striking that the frequency of the brackets increases significantly on a temporal axis (its frequency quadruples from under 5 occurrences per 1,000 tokens to nearly 20). This is especially noteworthy

because the frequency of most other punctuation units in the GraphVar corpus tends to decrease. In recent years, many final exams, for example, are written without semicolons, dashes, apostrophes, colons, question marks, and exclamation marks (cf. Berg & Romstadt 2021: 2017, Romstadt in preparation). While other punctuation units are becoming more and more marginal in usage, this is not the case for brackets. Assuming that punctuation units are only used when they are functional, this suggests from a broader perspective that the function of the brackets seems particularly and increasingly relevant in the communicative context outlined.

Additionally, it is striking that the final bracket <)> is always (!) slightly more frequent than the initial one. This seemingly contradicts the assumption that the closing bracket is more likely to be forgotten than the opening one (especially in longer constructions within the brackets). The reason for the higher frequency of the closing symbol is brackets like (7).

- (7) 1.) (German exam 2018, grade D+)
1.)

Brackets like (7) have the main function to organize and structure the text; they primarily serve a text-structural role. Baudusch (1989: 171) calls them “half brackets.” Referring to their main function—textual organization—they are referred to here as *organization brackets*.

Figure 4 shows the global frequency of bracket forms from a macro perspective. It is possible that a few exams contain an unusually large number of brackets, which could distort the data in favor of brackets. Another measure for bracket usage is therefore prevalence. This is a metric commonly used in epidemiology and medical statistics to quantify the distribution of diseases. In this case, it counts how many exams contain at least one bracket. An exam with one bracket is treated the same as one containing 100 brackets in the analysis.

In Figure 5, it is evident that in the early years, a large number of exams contain at least one bracket. When combining this observation with the very low global frequency shown in Figure 4 for these periods, it becomes clear that there appear to be relatively many exams with only a few brackets. After this period, the global frequency of brackets remains relatively stable at a low level, while prevalence decreases. This means that there are fewer exams containing brackets, but those that do use them more frequently (the total number of brackets is distributed over fewer exams). Both developments (constant frequency and decreasing prevalence) are somewhat opposed to each other. From the 1960s onward, the observed increase in frequency becomes apparent in the prevalence as well. More and more brackets are being used in final exams—and in more and more exams. In the last decade, almost every final exam has contained at least one bracket.

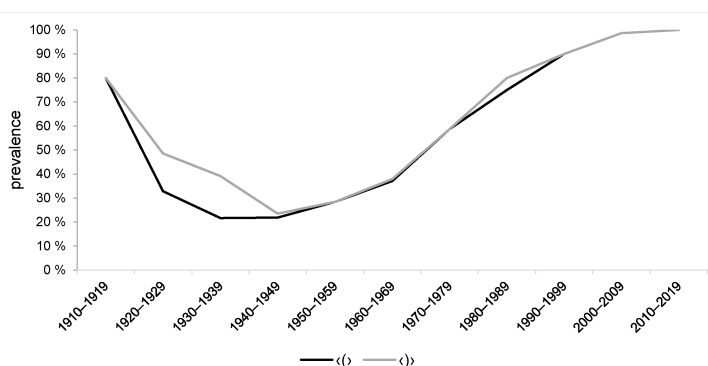


FIG. 5. Prevalence of round brackets over time

4.1. Form(s)

Up to this point, we have only spoken generally of “brackets,” whereas in the GraphVar corpus, only the following forms were actually analyzed: $\langle \rangle$ and $\langle \rangle$. However, brackets also appear in other forms in German. Four form variants are commonly distinguished: round, square, angle, and curly brackets. They can be described as allographic variants (cf. Meletis 2020b). Gallmann (1985: 171) proposes the following hierarchy of markedness.

less marked		more strongly marked	
$()$		$[]$	$\langle \rangle$
			$\{ \}$

FIG. 6. Hierarchy of markedness of allographic bracket variants (cf. Gallmann 1985: 171)

This exactly corresponds to the frequency of the individual bracket forms in the GraphVar corpus (see Table 2). At the same time, Table 2 also clearly shows that the round variant dominates.

The evaluations in Figures 4 and 5 have thus indeed depicted the formal prototype, which is more than 230 times more frequent than all other forms combined. This may also be due to the fact that the GraphVar corpus contains handwritten texts.

The frequencies of the other allographic bracket forms in the GraphVar corpus are too low to allow for reliable quantitative statements based on this data. For contextualization, one can consult the German Ref-

TABLE 2. Frequency of allographic bracket variants

Form	Frequency (absolute)	Frequency per 1,000 tokens
()	17,496	12.8
[]	73	0.0536
< >	2	0.00147
{ }	1	0.000734

TABLE 3. Frequency of allographic bracket variants in the corpus DERECHAR-v-uni-204-a-c-2018-02-28-1.0

7				
Form	Name	Unicode	Frequency (absolute)	Frequency (per 1,000 tokens)
(	left parenthesis	U+0028	180,360,930	1.00
)	right parenthesis	U+0029	179,601,830	0.999
]	right square bracket	U+005D	4,022,239	0.0224
[	left square bracket	U+005B	4,010,322	0.00223
{	left curly bracket	U+007B	413,060	0.000229
}	right curly bracket	U+007D	409,724	0.000228
>	single right-pointing angle quotation mark	U+203A	3,490	0.00000194
<	single left-pointing angle quotation mark	U+2039	3,008	0.00000167

erence Corpus (*Deutsches Referenzkorpus*, DEREKO). This is a large corpus with over 175 billion tokens, primarily containing press texts. In the DERECHAR corpus version, the frequency of Unicode characters is counted within that corpus (cf. Institut für Deutsche Sprache 2018). The result is shown in Table 3.

Since the individual Unicode characters are distinguished, a more precise differentiation between types of brackets can be made. It is important to note that the “angle brackets” in Gallmann’s system are described as quotation marks in DERECHAR. This does not change their overall very low frequency, which is also significantly lower than that of curly brackets, which were more frequent in the GraphVar corpus than the angle brackets.

Two other observations stand out: First, the initial (round) bracket is more frequent than the final one, what is the opposite in the GraphVar corpus across all years. This is ultimately a reflection of the texts included in the corpus. In longer, argumentative compositions, the need for textual structuring (including organizational brackets) is likely more

7. Cf. <https://www.ids-mannheim.de/fileadmin/kl/derewo/DeReChar-v-uni-204-a-c-2018-02-28-1.0.html> (accessed: 03.01.2025).

prominent than in prototypical press texts. It can be assumed that the final bracket is more often forgotten than the initial one. These two developments run in opposite directions. In newspaper texts, the initial bracket predominates, whereas in Abitur exams, the final bracket (as an organizational bracket) is more common.

Second, both corpora are homogeneous in that the round bracket form, which historically was a forerunner, remains the clearly dominant one. It is formally the prototype.

4.2. Function(s)

For the following functional analysis, the focus on brackets is methodologically narrowed. While in Section 4.1 both parts of the same punctuation unit “bracket” were counted separately, here instances with both initial and final brackets are counted together as a single token (8). Also included are all instances of “half brackets,” in which a bracket is essentially missing. These are, on the one hand, organization brackets (i.e., the true half brackets) like (9), and on the other hand, cases where a bracket was forgotten during writing, as in (10).⁸

- (8) Durch den hundertjährigen Krieg (1339-1453) war Frankreich in die größte Notlage geraten. (German exam 1920, grade C+)
The Hundred Years' War (1339-1453) had plunged France into its greatest crisis.
- (9) 1.) Abenteuerlust und Wandertrieb. (German exam 1944, grade D+)
1.) Thirst for adventure and wanderlust.
- (10) Sie fühlen sich von der Stadt abhängig (Monika verdient dort ihr Geld, dort gibt es Dinge zu kaufen, die es auf dem Lande nicht gibt. (German exam 1974, grade D)
They feel dependent on the city (Monika earns her money there, and there are things available to buy that cannot be found in the countryside.

(8)-(10) each corresponds exactly to one example in the following functional analysis. In total, 9,239 tokens are included in the analysis.

These examples are first described based on whether they are construction brackets or commenting brackets. These are the categories

8. It should be noted that this is, of course, only an interpretation of these examples. From a product perspective, it is not reconstructable why exactly one bracket was realized and why another was not. However, it can be assumed that final exams are generally produced with a focus on orthographic correctness. As stated in the educational standards: “Serious and repeated violations of linguistic accuracy [...] lead to a deduction of up to two points” (Kultusministerkonferenz 2012: 23). Therefore, it can be assumed that the most plausible interpretation is that in examples like (10), a bracket was actually forgotten. To empirically substantiate this, data on the writing process are needed, and exploring this has remained a research desideratum with few exceptions (e.g., Schüler 2023).

proposed by Bredel (2008), which are primarily based on a syntactic description of the relationship between the written units inside and outside the brackets. Construction brackets can be omitted without creating an ungrammatical structure. This is not the case with commenting brackets. A special case occurs when entire graphematic sentences are enclosed in brackets, e.g., (10).

- (10) (Wo liegen diese Gefahren und wie kann man ihnen begegnen?) (German exam 1950, grade B)
(Where do these dangers lie, and how can they be addressed?)

Since they form their own—complete—syntactic structure, they are primarily *textually* linked to the surrounding written material. Therefore, they are not simply classified as construction or commenting brackets but are counted as their own category. They are extremely rare in all years, so this is more of a side note (they occur a total of 184 times). Also counted separately are enclosed punctuation units, such as in (4), which are very rare as well (8 occurrences). Lastly, another category is added: organization brackets. These are primarily examples like (9), where the (half) brackets clearly serve a text-structuring function. They cannot easily be classified into the categories of construction and commenting brackets, which are mainly described based on the reference unit of the (syntactic) sentence.

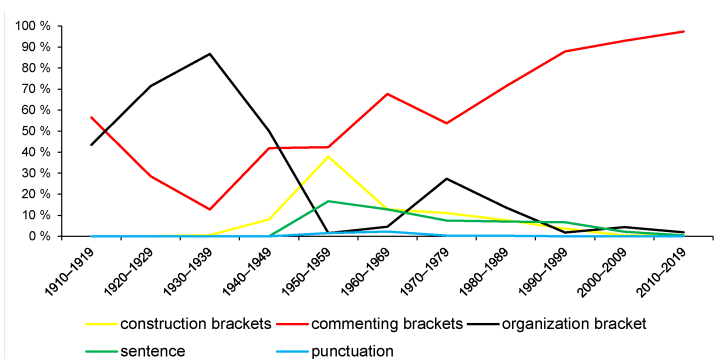


FIG. 7. Proportions of different structural bracket types in the GraphVar corpus

Three aspects are particularly noteworthy in Figure 7.

First, it is apparent that two categories of bracket usage are especially relevant: commenting brackets and organization brackets. Other configurations, in contrast, occur only to a negligible extent.

Second, a prototype emerges along a temporal axis. Brackets in German (in school leaving exams) today are almost always comment brackets. These are the instances that fit particularly well with the function

of the symbol, which is to establish a—also syntactic—subordinate discourse. In contrast, the organization bracket dominated much more strongly in the early decades. This can also be explained by structural requirements for the final exams. Until the 1970s, it was mandatory for students to include a writing plan for their exams. Planning processes were thus much more (literally) visible than they are today—and this is directly connected to a greater need for textual organization.

Third, the unusually high proportion of construction brackets in the 1950s is striking. During this time, there are particularly many adjective attributes and adverbials that are enclosed in brackets (and syntactically integrable), such as (11).

- (11) Arbeiten, die für den Menschen mühsam und zeitraubend waren, führt (heute) nun die Maschine aus, überwältigt sie in einem viel schnelleren Maße. (German exam 1952, grade C)

Tasks that were arduous and time-consuming for humans are (now) carried out by machines, which accomplish them at a much faster pace.

It is possible that the increased use of construction brackets here is a reflection of the topics being addressed. This could only be clarified through studies based on broader empirical data. However, it remains constant that such bracket usages never make up the quantitatively most significant part. These are always organization and commenting brackets.

Regarding the latter category, there are also brackets that stand out, which are used in the context of citations. This concerns, for example, the indication of sources directly after quoted units.

- (12) Diese Geliebte „erquickt“ (V. 18) das lyrische Ich mit „ew[i]ger Treue“ (V. 17). (German exam 2018, grade B)

This beloved “revives” (v. 18) the lyrical subject with “eternal loyalty” (v. 17).

When evaluating the proportion of such *citation brackets* among all commenting brackets, it becomes clear that the development in Figure 7 is primarily due to these instances.

For contextualization, reference can be made to the study by Sieber (1998), who also examined German school leaving exams over a temporal scale. His corpus includes 112 texts, which were written at a single school between 1881 and 1991. He notes: “It does not seem like much has changed in a significant way” (Sieber 1998: 112). However, on a global level, he observes a “convergence of written and spoken language, with the trend clearly moving towards spoken language” (Sieber 1998: 51).

He describes this with the term “Parlando.” The main thesis of the text is that school leaving exams are increasingly becoming such *Parlando* texts. They are characterized by, among other aspects, not adhering uniformly to linguistic norms, which is, e.g., reflected in an “extraordinarily careless handling [...] of the spelling of individual words [...] and

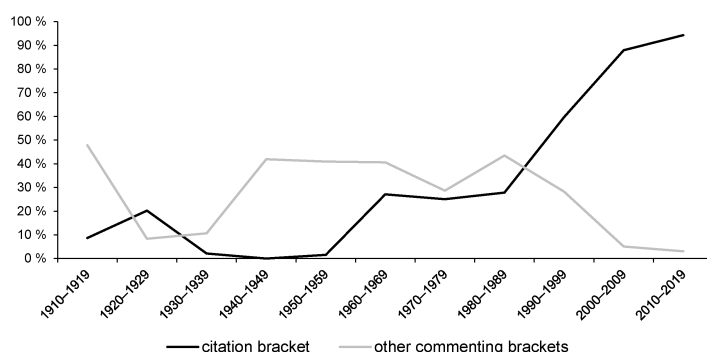


FIG. 8. Proportion of citation brackets and other commenting brackets among all bracket instances in the GraphVar corpus

the placing of punctuation units [...].” Interesting remarks are made regarding citations: “While the early years contain several texts with citations, they almost entirely disappear in recent years” (Sieber 1998: 113). He thinks it’s “remarkable that despite the demands posed by the examination situation, in more recent times much more personalized and freer forms have become possible [...] whereas older texts almost completely lack such a perspective” (Sieber 1998: 131, 148).

This contradicts the observation made in the GraphVar corpus: Here, it is precisely the *increasing* relevance of citations that has led to the overall rise in the frequency of the brackets. Moreover, this is not a development that has only occurred in the last two decades, i.e., after Sieber’s study, but has been visible since at least the 1960s.

Furthermore, it can be contextualized with the changing demands on school leaving exams, which are anticipated by the writers. While, particularly after World War II, “personalized” forms in reflection essays were much more common, this is no longer the case today.⁹ More or less scientific discussions on a professional level are expected nowadays.

The educational standards in the subject of German, therefore, consistently refer to “text-related” or “material-supported” writing tasks (cf. Kultusministerkonferenz 2012: 24). In both cases, the reference to external texts is already clear terminologically. It is either about *text*-based analyses/discussions or the processing of other sources/*material* in one’s own texts. What both task types have in common is the positioning of the writing individual in relation to other written statements, which can

9. As stated in the educational standards: “Creative writing’ [...] should primarily be reserved for classroom work and is excluded as the sole task in the written Abitur exam. [...] The text prompt must [...] not serve merely as a trigger for subjective or imitative writing” (Kultusministerkonferenz 2012: 24).

understandably be realized primarily through citations. “Text reference” as a quality feature of such exams is then merely a “disguised” demand to substantiate arguments with citations. This becomes evident in Figure 4, even though the bracket as a punctuation unit is neither formally nor functionally tied to citations.

Up to this point, the so-called functional analyses were mainly based on structural aspects (cf. the terms construction and commenting brackets). It was primarily about the relationship (and syntactic integrability) of the units within the brackets in relation to the surrounding co-text.

Only the evaluation of citation brackets was already functionally motivated in the narrower sense.¹⁰

To place this perspective more centrally in the following, the categories proposed by Levin & Ström Herold (2021) will also be applied to the data discussed here (cf. Section 2). Only the organization brackets, which cannot easily be assigned to another subclass, are described as their own category. First, the relationship between content-oriented and interpersonal brackets is analyzed.

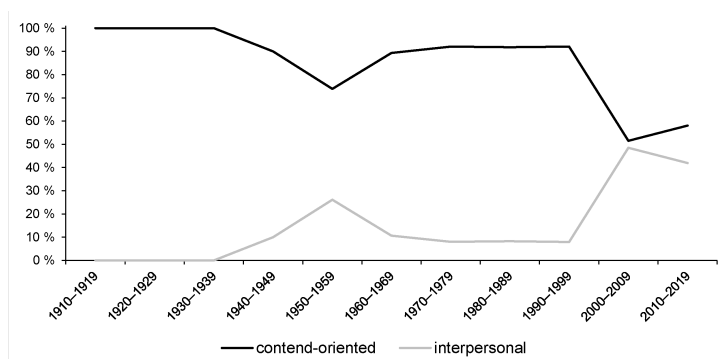


FIG. 9. Percentage of content-oriented and interpersonal brackets in the Graph-Var corpus

From a macro perspective, it becomes clear that brackets in Abitur exams are primarily content-oriented. It is only in the most recent years that the relevance of interpersonal brackets has increased. This is most comparable to Sieber’s (1998) observations about Parlando texts. Generally, it is nonetheless not surprising that school leaving exams are more content-oriented. Ideally, they are subject-specific statements that (among other things) are intended to demonstrate the competence of

10. Compare Rödel & Rothstein (2015: 235): “The function of a linguistic sign results from its meaning and its use.”

the writer in answering a (complex) question in a way that earns them the school graduation.

At the same time, one could also interpret the data from an opposing perspective: brackets offer an (accepted) way of expressing oneself subjectively or more generally in an interpersonal manner. For example, if a writer wants to express a subjective author comment in a school leaving exam, it is better to do so within brackets than without this punctuation unit. However, this clearly has not been done to a quantitatively significant extent—at least not until the 2000s.

The data from Figure 9 can be further differentiated when the subtypes of interpersonal and content-oriented brackets are considered. This is done in Figure 10. The organization brackets, which were discussed in more detail above, are included as a separate category.

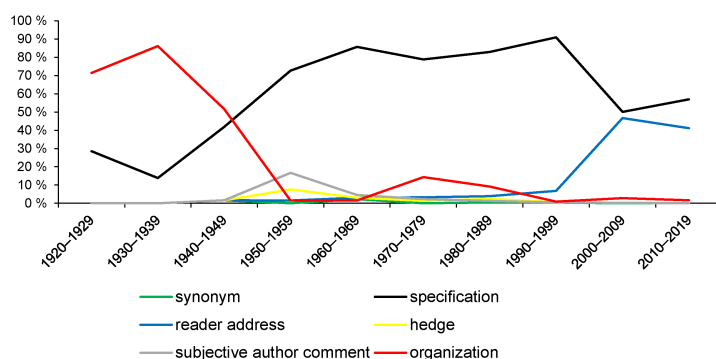


FIG. 10. Proportions of Functional Bracket Types According to Levin & Ström Herold (2021) in the GraphVar Corpus

It is initially noticeable that the corpus primarily contains specification brackets. The other content-oriented subcategory, the indication of synonyms, plays no role in German school leaving exams.¹¹ Thus, content specification brackets, which provide additional information to the surrounding co-text, are particularly relevant. On the level of interpersonal brackets, hedge and subjective author comments, conversely, make up a small proportion. In both cases, it concerns the subjective commentary on the context, which is generally not expected in exams. This is different for the category of reader address, which has notably increased

11. Here, the dependency of punctuation usage on the subject becomes clear. In the biology exams, which are also included in the GraphVar corpus but were excluded from this study, synonym brackets appear to a greater extent, mainly due to the higher density of foreign words and technical terms in those papers (cf. Romstadt in preparation).

in relevance, especially in the last two decades. When writers of school leaving exams reveal themselves interpersonally in brackets, it is primarily about the readers. This is also well-founded: the (anticipated) readers of those texts are the ones who evaluate them. Their judgment matters significantly (in sociology, this is referred to as “high-stakes tests”). Accordingly, it is not surprising that such instances are the entry point for interpersonal brackets in school leaving exams. Furthermore, it is often not just any reader address that appears in brackets, but it can be linked to what was noted above about citations.

- (13) (Dieses werde ich später erklären) (German exam 1974, grade C)

(I will explain this later)

- (14) Botho sagt deutlich, dass sie für ihn kein Glück bedeutet (vgl. Z. 40–41). (German exam 2003, grade B-)

Botho clearly states that she does not bring him happiness (see lines 40–41).

Examples like (13) are the absolute exception. Here, the student indicates within the brackets that the explanation anticipated by the readers at this point will follow later in the text (meaning it is not forgotten, but only postponed). The prototype of reader address brackets in school leaving exams, on the other hand, is unquestionably (14). Nearly all examples refer to constructions with <vgl.> and <s.> or <siehe>. Both are reader address markers because they can be interpreted as imperative verb forms (“vgl. = Compare!” cf. Duden 2020: 1211; “s. = See!” cf. Duden 2020: 973—both with exclamation marks). If these forms are understood as abbreviated imperative forms, they are much more likely to be reader addresses than the possible specification instances for enriching the text with information about the location of the cited units.

Taking these functional observations together, it becomes clear how the GraphVar corpus differs from that of Levin & Ström Herold (2021). While the latter contains a greater variety of different bracket functions, school leaving exams primarily feature two key subtypes. One of these is content-oriented specification brackets, whose proportion is roughly the same as in Ström-Herold & Levin (2021), which is a notable insight as well. Both corpora do not differ significantly in this regard ($\chi^2 = 1.996, df = 1, p = 0.1577$). On the other hand, almost only reader address brackets appear, and they are significantly and statistically more frequent than in the corpus of Levin & Ström Herold (2021) ($\chi^2 = 111.15, df = 1, p < 2.2 \times 10^{-16}, \phi = 0.11$).

This results in a functional prototype in addition to a formal one in the GraphVar corpus.

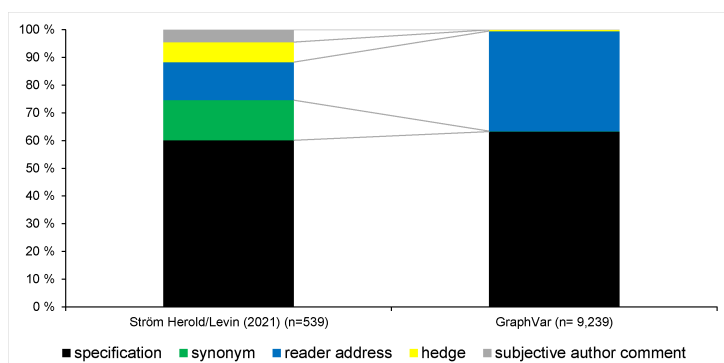


FIG. 11. Comparison of the proportions of different bracket functions in the GraphVar corpus and in Levin & Ström Herold (2021)

5. Beyond the Horizon: Brackets and the Plus Sign(s) as Kindred Spirits

In several studies on German punctuation, brackets are compared with the (paired) dash. Fuhrhop & Schreiber (2015: 48), for instance, categorize these two units, along with the comma, as tools for marking parentheses in written German. They regard brackets as stronger markers of parentheses compared to dashes, noting that while both units independently convey additional pragmatic information, brackets are inherently paired.

Ström Herold & Levin (2024) contrast their findings on brackets with those on dashes, motivated by functional overlaps between the two units—specifically their shared role in marking parentheses within a sentence (cf. Ström Herold & Levin 2024: 263–264).

However, Bredel (2008: 146) identifies a key distinction: content within brackets is seen as a deliberate authorial disclosure, while the use of dashes often stems from unsatisfactory planning during ongoing verbalization on the speaker's part.

Ström Herold & Levin (2024: 264) observe that “dashes [are] having a more informal, dramatic flair than both brackets and commas.”

While these observations are valid, they pertain to different levels of the linguistic system (e.g., pragmatics, obligatory pairing as a formal property, and syntactic-structural features of the marked units).

In this chapter, we will instead focus on a punctuation mark that, in specific contexts, better fulfills the claim of complementarity to brackets—especially in terms of many of the aforementioned aspects: the plus sign <+>.

5.1. The Plus Sign and the German Punctuation System

Admittedly, the plus sign, unlike the paired dash,¹² does not appear in the literature on (German) punctuation. However, there are specific contexts in which a justified argument can be made for considering the plus sign as part of the German punctuation system. These contexts are particularly prevalent in German online journalism. To mark breaking news, live blogs, or to indicate a new development in an ongoing or past event, plus signs are a common tool employed by online editorial teams. Typically, at least two, and often three, plus signs are placed both before and after the message (see Figure 12). These signs frame the sequence of words deemed most interesting or important for the reader. By using these symbols, journalists aim to convey immediacy, relevance, or both simultaneously.

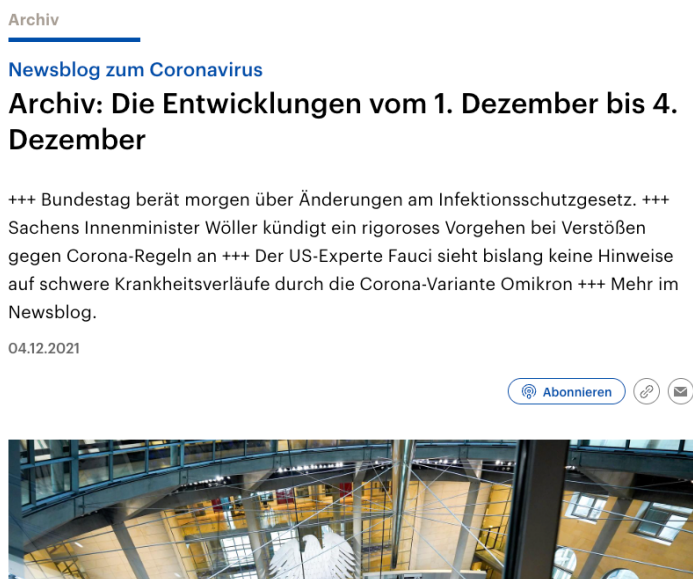


FIG. 12. Newsblog coronavirus, <https://www.deutschlandfunk.de/newsblog-zum-coronavirus-114.html> (accessed: 09.09.2024)¹³

12. See, for example, Bredel (2008: 23–24), Baudusch (1981: 207), and Mentrup (1983: 34–39).

13. Translation: “News blog on the coronavirus. Archive: Developments from December 1st to December 4th. +++ The Bundestag will discuss changes to the Infection Protection Act tomorrow. +++ Saxony’s Minister of the Interior Wöllner announces a

For recipients, these symbols serve an additional function in this context: they facilitate reading. By employing plus signs, a text can be segmented, signaling a thematic shift to the reader. As illustrated in Figure 12, such markers separate individual reports from one another. Through this structuring function, the symbols meet a key criterion for classification as punctuation units: “The purpose of punctuation units is to structure a written or printed text clearly and thereby ensure its comprehensibility and readability” (Baudusch 1981: 207). The implementation of the segmenting function is particularly evident in examples where, for improved clarity, different pieces of information on the same topic or “individual news items are demarcated with +++” (Meier-Vieracker 2021).

+++ 18:12 Brasiliens Gesundheitsminister positiv auf Corona getestet +++

Brasiliens Gesundheitsminister Eduardo Pazuello hat sich mit dem Coronavirus angesteckt. Pazuello sei positiv auf das Virus getestet worden, teilt sein Ministerium mit. Er ist bereits der dritte Gesundheitsminister des südamerikanischen Landes seit Beginn der Coronavirus-Pandemie.

+++ 17:37 In der Lombardei startet nächtliches Ausgangsverbot +++

Die italienische Region Lombardei hat wegen der stark steigenden Corona-Zahlen nächtliche Ausgangsverbote für die rund zehn Millionen Bürger erlassen. Die verschärften Regeln in der norditalienischen Region, zu der auch Mailand und Bergamo gehören, gelten nach dem Erlass ab Donnerstagabend. Die Menschen dürfen dann zwischen 23 Uhr und 5 Uhr ihr Haus nur noch aus wichtigem Grund wie Arbeit oder Krankheit verlassen. Nach Angaben der Zeitung “Corriere della Sera” können die Behörden bei Verstößen Strafen zwischen 400 und 3000 Euro verhängen.

+++ 17:10 Gesundheitsminister Spahn positiv auf Corona getestet +++



Kabinett nicht in Quarantäne
Jens Spahn positiv auf Corona getestet

+++ 16:58 EU will Kanada von Liste sicherer Staaten streichen +++

Die EU will Kanada, Georgien und Tunesien von ihrer Liste jener Länder streichen, aus denen Einreisen ohne Auflagen erlaubt sind. Bei einem EU-Botschaftertreffen wird als Grund ein starker Anstieg der Infektionen genannt. Sollte kein EU-Land Einwände erheben, wird die Liste bis Ende der Woche entsprechend aktualisiert, wie zwei EU-Diplomaten sagen. Weil sich die Lage in Singapur entspannt

FIG. 13. Liveticker Coronavirus, <https://www.n-tv.de/panorama/23-55-Mehr-als-220-000-Tote-in-den-USA-article22116061.html> (accessed: 14.08.2024)¹⁴

strict approach to violations of Corona rules. +++ US expert Fauci sees no evidence of severe disease courses caused by the Corona variant Omicron so far. +++”.

14. Translation of the headings including plus signs: “+++ 18:12 Brazil’s Health Minister tests positive for Corona +++ [...] +++ 17:37 Nighttime curfew begins in Lombardy

In Section 2, we already addressed the characteristics of punctuation units as defined by Bredel (2008; 2020). To compare the plus sign with brackets, it must first be examined whether it can be classified as a punctuation unit, defined as “representable, non-recodable, non-combinable, single-element segments of writing” (Bredel 2020: 9). Since the plus sign, like letters, numerals, and special characters, belongs to the smallest units of writing its inclusion among segmental tools requires no further debate.

The criterion of single-element nature is also quickly addressed. Unlike letters, which can be realized in uppercase and lowercase forms, the plus sign exists in only one form. It also cannot be combined with other graphic elements, such as diacritics (cf. Günther 1988: 64–65), and is thus, like punctuation units, non-additive. Additionally, the feature of standalone identifiability is evident through the sign’s realization: <+>.

In addition to that, combining multiple plus signs does not result in a larger unit. Even when repeated, each plus sign remains an individual graphic element, unlike digits, which, when placed together, form a larger numerical entity (cf. Günther 1988: 67–68).

Thus, the plus sign satisfies almost all criteria for punctuation units unequivocally. The property of non-recodability, however, remains debatable. In certain contexts, the plus sign can represent *and* or *plus*, making it recordable—which would mean it’s more of a special character than a punctuation unit. Consider the following sentence to be read aloud:

- (15) Nur so ist auch das Skandalon des monatelangen Zögerns der damaligen Bundesregierung zu erklären, in den 2+4-Verhandlungen die Oder-Neiße-Grenze anzuerkennen – obwohl die Deutschen gegenüber Polen, dem viermal geteilten, brutal überfallenen und versklavten Nachbarn im Osten, eine mit nur wenig sonst zu vergleichende historische Schuld tragen. (Fischer 2000: 5)

This is the only way to explain the scandal of the months-long hesitation by the then German government during the 2+4 negotiations to recognize the Oder-Neisse border—despite the fact that the Germans bear a historical guilt toward Poland, their eastern neighbor, which was partitioned four times, brutally attacked, and enslaved, that is hardly comparable to any other.

The verbalizability of the plus sign (as [plʊs] in German) is undeniable. According to Bredel’s perspective, this would mean the sign could no longer be included in the inventory of the German punctuation system. Bredel (2020: 9) explicitly tightens the criterion for recodability, stating: “To qualify as recodable, it is sufficient to identify a context where this characteristic applies.” This criterion leads her to exclude symbols such as the asterisk from the category of punctuation units, as it can stand for

+++ [...] +++ 17:10 Health Minister Spahn tests positive for Corona +++ [...] +++ 16:58 EU to remove Canada from the list of safe countries +++”.

born in (cf. Bredel 2008: 24). She would likely argue similarly regarding the plus sign.

However, Bredel's inventory also includes units such as the hyphen and the colon (by Bredel as well as by many other linguists¹⁵). Yet, there are also contexts in these units that seem to be recodable:

- (16) Seit Freitag wird auf der Bahnstrecke Berlin-Hamburg groß gebaut.
(Haase-Wendt 2024)
Since Friday, major construction work has been underway on the Berlin-Hamburg railway line.

That the hyphen here stands for *to* or *through* and would therefore be pronounced is plausible. This becomes even more evident in the following example sentence:

- (17) Die Baskets unterlagen dem italienischen Topteam Pallacanestro Reggiana mit 91:94 und zeigten über 40 Minuten zwei unterschiedliche Gesichter.
(Halbfas 2025)
The Baskets were defeated by the Italian top team Pallacanestro Reggiana with a score of 91:94, displaying two completely different sides over the course of 40 minutes.

The recodability of the colon in (17) is obvious. These examples illustrate the problems that arise with a strict criterion of recodability. If even units whose inclusion in the inventory is widely agreed upon do not meet all the listed criteria in every context, then the discussion of the plus sign as part of the German punctuation system is justified. While there are indeed contexts in which the plus sign is likely to be verbalized, this does not automatically exclude the sign itself from the inventory debate. There are examples where the plus sign is (probably) not verbalized and otherwise fulfills Bredel's criteria.

When the graphical element appears as a structuring device, as analyzed above, and cannot be verbalized, this phenomenon will hereafter be referred to as the "liveticker plus sign."

5.2. Comparison Between Brackets and Plus Sign

Bredel (2008: 146–147) conducts a comparison between brackets and paired dashes, despite the latter not being classified as a distinct punctuation unit in her framework. While the final element in brackets is obligatory, the realization of a second dash is considered a "specific usage variant" (Bredel 2008: 23–24).

15. Schreiber (2020: 7) provides an overview of the inventories from various studies on German punctuation. Symbols such as the period, the hyphen and colon are included in all of them.

The situation differs for the live-ticker plus sign, which never appears alone but always as an initial and final block. No evidence suggests that this symbol is used only before or only after the text. Therefore, we can conclude that the final element is mandatory, justifying the inclusion of the live-ticker plus sign in the inventory of paired units. However, Bredel emphasizes that brackets and quotation marks are the “only truly paired symbols” (Bredel 2008: 82), distinguished by differing appearances of their initial and final elements (cf. Bredel 2020: 10). Since the plus signs on either side of a text are visually identical, one might argue against their classification as “truly paired,” challenging their inclusion in the German punctuation system.

Despite this, there are sufficient contexts where the plus sign contrasts effectively with brackets. Unlike the dash, the live-ticker plus sign requires the realization of the final element. This formal necessity aligns more closely with brackets than with dashes. Examples (18) and (19), repeated from Figures 12 and 13, illustrate this distinction. In (19), the plus signs are obligatorily paired, while in Figure 12, the “initial” and the “final” plus signs may appear intermixed within a paragraph.

- (18) +++ Bundestag berät morgen über Änderungen am Infektionsschutzgesetz.
 +++ Sachsens Innenminister Wöllner kündigt ein rigoroses Vorgehen bei
 Verstößen gegen Corona-Regeln an +++ Der US-Experte Fauci sieht bis-
 lang keine Hinweise auf schwere Krankheitsverläufe durch die
 Corona-Variante Omikron +++
 +++ *The Bundestag will discuss changes to the Infection Protection Act tomorrow.*
 +++ *Saxony's Minister of the Interior Wöllner announces a strict approach to vio-*
lations of the Corona rules. +++ US expert Fauci sees no evidence of severe disease
courses caused by the Corona variant Omicron so far. +++
- (19) +++ 18:12 Brasiliens Gesundheitsminister positiv auf Corona getestet +++
 [...] +++ 17:37 In der Lombardei startet nächtliches Ausgangsverbot +++
 [...] +++ 17:10 Gesundheitsminister Spahn positiv auf Corona getestet +++
 +++ 18:12 Brazil's Health Minister tests positive for Corona +++
 [...] +++ 17:37 Nighttime curfew begins in Lombardy +++
 [...] +++ 17:10 Health Minister Spahn tests positive for Corona +++

The plus sign's pairing is evident in overlapping contexts, as initial and final elements of the whole sequence are consistently realized.

Thus, live-ticker plus signs exhibit a stronger pairing characteristic than dashes. Both symbols—brackets and plus signs—serve as framing devices for text sequences. Brackets typically indicate an aside (see Section 2), which is why they are often compared with dashes. However, the live-ticker plus sign serves a contrasting purpose, directing the reader's attention to the text's most important information.

Whereas brackets therefore function as the “strongest symbol for asides” (Fuhrhop & Schreiber 2015: 48), the live-ticker plus sign expresses emphasis or urgency, akin to a carnival barker shouting, “Don't

miss this! This is important!” (Niggemeier 2014). While brackets highlight *secondary* remarks, plus signs emphasize the author’s *most significant* statement.

Though Fuhrhop & Schreiber (2015: 49) note that dashes also emphasize certain elements, and Lauchman (2010: 114–115) considers dashes the most expressive punctuation, the live-ticker plus sign often surpasses the dash in expressiveness. For example, in public transport displays, plus signs highlight disruptions or urgent updates, signaling critical information to passengers. This attention-grabbing function is unique, even compared to exclamation points, which lack the plus sign’s pairing feature.

Thus, the plus sign is the punctuation unit most similar to brackets in formal terms, particularly because of its pairing. Although quotation marks are also “truly paired,” their function differs fundamentally, as they distance the author from the enclosed text (cf. Klockow 1978: 15). Brackets and plus signs, by contrast, reflect the author’s voice, either adding supplementary information (brackets) or emphasizing a core statement (plus signs). These similarities underscore a formal-functional relationship between the two units, making them suitable for contrasting purposes.

However, the live-ticker plus sign remains primarily associated with specific communicative settings. Its name alone implies its role in such contexts, which explains why it has not been extensively discussed in punctuation studies. Bredel (2008) notes that her analysis of German punctuation focuses on “text mode,” characterized by a lack of (syntactic) structuring by line construction (Bredel 2008: 32). Reißig (2015) refines this by emphasizing the continuity of text mode, where elements could theoretically appear in an uninterrupted sequence without altering meaning or grammar (Reißig 2015: 33). This description aligns perfectly with live-ticker text, which unfolds in seemingly endless sequences, either visually or through simulated formatting (see Figure 12).

In such cases, plus signs resemble headings, serving as an introduction to subsequent content (see Figure 13). They function as “graphical distinctions in a spatial relationship to the text” (Finkbeiner 2021: 265). While headings are not confined to text or list mode, their defining feature is graphical separation from the main text. Live-ticker plus signs are a practical graphical distinction, especially when other typographic methods are unavailable.

Future studies could investigate the quantitative relationships between different uses of the plus sign, distinguishing between verbalized and non-verbalized contexts. Such research would provide an empirical basis for enriching the “recodability” criterion with “verbalizedness” and refining distinctions between core punctuation units, peripheral symbols, and transitional cases.

In conclusion, plus signs and brackets share notable similarities, particularly in their pairing and their functions of conveying authorial intent. While brackets contextualize the base text, plus signs emphasize its main point. Both segment text effectively, with brackets interrupting (sentence) linearity and plus signs separating updates in live-ticker formats. Unsurprisingly, plus signs are popular in online journalism, where concise, impactful communication is paramount.

Yet, differences remain. Brackets can integrate into any text type, while plus signs seem to be limited to specific contexts like online journalism and announcements. Although the inclusion of the plus sign in the German punctuation system remains unsettled, its role as a peripheral element merits consideration. Its text-segmenting function and fulfillment of key criteria in specific contexts clearly support this.

6. What Follows? Implications for Punctuation Research and Grapholinguistics

The aim of this study was the corpus-driven linguistic description of bracket usage in German. We argued that school leaving exams (*Abiturarbeiten*), particularly due to their temporal scope, are a suitable data source for this purpose.

We demonstrated that, formally, the round bracket is by far the dominant form in these handwritten texts. It appears 203 times more frequently than all other allographic variants combined. Structurally, most brackets are commenting brackets found within graphematic sentences. Particularly in recent years, they predominantly relate to citation, explaining why this punctuation unit—unlike nearly all other components of the inventory—has become increasingly common. Nowadays, nearly every graduation exam contains at least one bracket.

This was not the case in the past, as the structural prototype of the commenting bracket has, over the last century, supplanted the organization bracket, which serves a text-structuring function (and is often represented as a “half-bracket” in Baudusch’s (1989) sense).

Functionally, the prototype is mainly the specification bracket, which is thus described as content-oriented. This is well justified within the communicative context of school leaving exams, although other linguistic studies have drawn differing conclusions (cf. Sieber 1998).

The increasing use of interpersonal brackets in recent decades is also due to the growing importance of citation, as these brackets often contain imperative forms like <vgl.> or <s.> (*compare!* or *see!*). These are interpersonal as the writers instruct readers to take a specific action. Interpersonal brackets are one-way communicative tools in this regard, as the subjective commentary function for the surrounding text is virtually non-existent. All of this aligns well with what has been expected from

school leaving exams, particularly in recent years, as outlined in educational standards.

Thus, usage-based data reveals a prototype for German school leaving exams of a round, structurally commentary, functionally specifying bracket.

Building on this, we examined the liveticker plus sign, a unit that has been rarely (if ever) discussed in the context of punctuation. It was compared with this prototype. In certain communicative settings, specifically live-ticker scenarios, they can be described as non-recodable (punctuation) units. In this usage, they function as mirror images of brackets. Graphotactically, they are analogous, as they also appear in pairs. Functionally, however, they do not situate the enclosed written units within a side discourse but instead highlight them as particularly relevant. This usage unquestionably belongs to the text mode.

Further quantitative studies are needed to determine the general relevance of this phenomenon and the functional contrast with brackets. Such data could simultaneously contribute to refining the criterion of recodability. This would also impact other borderline cases of unit classification, such as the slash or the asterisk (even in the context of gender-neutral writing in German). Taking this further, it could ultimately lead to rethinking the category of “punctuation unit” not as a fixed concept but as a continuum. Prototypical punctuation units like commas or periods would then, based on quantitative usage observations, be situated alongside classic special characters like the <%>-sign and units in transitional areas. The liveticker plus sign, for instance, could be described as more punctuation-like than the equals sign, for example, the slash as more punctuation-like than the liveticker plus sign, and the colon (since it is verbalized in only a few contexts) as more punctuation-like than the others.

Thus, this study advocates for increased attention to usage-based studies in punctuation and grapholinguistics overall. This was demonstrated on multiple levels here. The relevance of the contrast between brackets and plus signs was described using usage data. The formal and functional characterization of brackets in German was also based on authentic instances, consistent with concepts like usage-based graphematics (UxG). This perspective necessarily connects production data (as in this study) with reception-based insights. Thus, analyses like those conducted here are more of a starting point than the culmination of grapholinguistic work. They are closely tied to other areas of this sub-discipline, including:

- corpus-driven studies: Which variations appear where? Which variants are used in which contexts and why? What does it mean if the liveticker plus sign predominantly appears in press contexts, where it contrasts with brackets? How can developments in bracket usage in

high school exams be explained in consideration of other punctuation units in the inventory?

- the system of punctuation units: Which functions (and forms) of punctuation units represent the core and which belong in the peripheral area?
- typological research: How can differences across writing systems be categorized? Brackets are—like punctuation units in general—a success story in writing. Meletis & Romstadt (accepted) compare punctuation inventories across 30 writing systems worldwide. Brackets appeared in the same form and with the same function as in German in 24 systems (80%). In two writing systems, the function of brackets is taken over by a graphically similar symbol, namely <||> in Javanese and <『』> in vertically written Korean Hangul. Only Barnun, Bassa, Buginese, and Balinese do not have brackets.
- grapholinguistics as a whole: What role do punctuation units play in writing and reading? How do they relate to standard concepts such as the grapheme? What historical developments can be traced?

Usage studies are the first necessary step to address these (and similar) questions from a grapholinguistic perspective.

References

- Baudusch, Renate (1981). “Die Prinzipien der deutschen Interpunktion”. In: *Zeitschrift für Germanistik* 2.2, pp. 206–218.
- (1989). *Punkt, Punkt, Komma, Strich. Regeln und Zweifelsfälle der deutschen Zeichensetzung. Regeln und Gebrauch der Satzzeichen mit Beispielen und Übersichten*. 3rd ed. Leipzig: Bibliographisches Institut.
- Baumgarten, Nicole, Bernd Meyer, and Demet Özçetin (2008). “Explicitness in translation and interpreting: A critical review and some empirical evidence (of an elusive concept)”. In: *Across Languages and Cultures* 9.2, pp. 177–203.
- Behrens, Ulrike (1989). *Wenn nicht alle Zeichen trügen. Interpunktion als Markierung syntaktischer Konstruktionen*. Frankfurt am Main: Lang.
- Berg, Kristian and Jonas Romstadt (2021). “Reifeprüfung—Das Komma in Abituraufsätzen von 1948 bis heute”. In: *Die Sprache in den Schulen—Eine Sprache im Werden. Dritter Bericht zur Lage der deutschen Sprache*. Ed. by Deutsche Akademie für Sprache und Dichtung and Union der deutschen Akademien der Wissenschaften. Berlin: Schmidt, pp. 205–236.
- Berg, Kristian, Jonas Romstadt, and Cedrek Neitzert (2021). *GraphVar—Korpusaufbau und Annotation. Version 1.0*. Accessed: 05.01.2021. URL: <https://graphvar.uni-bonn.de/dokumentation.html>.

- Biber, Douglas and Bethany Gray (2016). *Grammatical Complexity in Academic English. Linguistic Change in Writing*. Cambridge: Cambridge University Press.
- Bölling, Rainer (2010). *Kleine Geschichte des Abiturs*. Paderborn: Schöningh.
- Bredel, Ursula (2008). *Die Interpunktion des Deutschen. Ein kompositionelles System zur Online-Steuerung des Lesens*. Tübingen: Niemeyer.
- (2020). *Interpunktion*. 2nd ed. Heidelberg: Winter.
- Bredel, Ursula and Angelika Wöllstein (2024). “Neue Konzepte und Zugänge zur Zeichensetzung im Amtlichen Regelwerk”. In: *Orthographie in Wissenschaft und Gesellschaft. Schriftsystem—Norm—Schreibgebrauch*. Ed. by Sabine Krome et al. Berlin/New York: De Gruyter, pp. 63–80.
- Busch, Florian (2021). *Digitale Schreibregister. Kontexte, Formen und metapragmatische Reflexionen*. Berlin/Boston: De Gruyter.
- Döblin, Alfred (1947). *Berlin Alexanderplatz. Die Geschichte vom Franz Biberkopf*. Originally published in 1929. Kassel: Harriet Schleber Verlag.
- Duden (2020). *Duden. Die deutsche Rechtschreibung*. 28th ed. Mannheim: Dudenverlag.
- Eiben-Zach, Britta (2023). “Reifeprüfungsaufsätze als Quellen historischer Forschung”. In: *Abitur und Abituraufsätze zwischen 1882 und 1972. Prüfungspraktiken, professionelle Debatten und Aufsatztexte*. Ed. by Michael Kämper-van den Boogaart et al. Bad Heilbrunn: Klinkhardt, pp. 132–147.
- Finkbeiner, Rita (2021). “Sprechakttheoretische Überlegungen zur Typographie—am Beispiel von Presseüberschriften”. In: *Zeitschrift für Germanistische Linguistik* 49.2, pp. 244–291.
- Fischer, Joschka (2000). *Außenpolitik im Widerspruch. Was Gustav Stresemann erreichte, woran er scheiterte und was daraus gelernt wurde*. Accessed: 14.01.2025. URL: <https://www.zeit.de/2000/06/200006.t-stresemann.xml>.
- Fuhrhop, Nanna and Niklas Schreiber (2015). “Hauptgleis — Nebengleis — Weiche. Wie wir Einschübe sichtbar machen können”. In: *Praxis Deutsch* 254, pp. 48–53.
- Fürstenberg, Maurice (2023). *Kommagebrauch im Deutschen. Eine empirische Untersuchung zur Kommasetzung beim Schreiben und Einsetzen*. Berlin/Boston: De Gruyter.
- Gallmann, Peter (1985). *Graphische Elemente der geschriebenen Sprache. Grundlagen für eine Reform der Orthographie*. Tübingen: Niemeyer.
- Gottsched, Johann Christoph (1766). *Kern der Deutschen Sprachkunst, aus der ausführlichen Sprachkunst Herrn Professor Gottscheds, zum Gebrauche der Jugend, von ihm selbst ins Kurze gezogen*. 5th ed. Leipzig: Breitkopf. URL: <https://digital.lb-oldenburg.de/vd18/content/titleinfo/686469>.
- Günther, Hartmut (1988). *Schriftliche Sprache. Strukturen geschriebener Wörter und ihre Verarbeitung beim Lesen*. Tübingen: Niemeyer.
- Haase-Wendt, Björn (2024). *Schienenersatzverkehr. Pendlerfrust auf der Strecke Berlin-Hamburg*. Accessed: 27.09.2024. URL: <https://www.rbb24.de/>

- panorama / beitrag / 2024 / 08 / berlin - hamburg - bahn - ersatzverkehr - baustelle - bus - prignitz . htm / answerTo = a2015260 - 2944 - 4659 - 952a - d32d51a87270/Twitter.
- Halbfas, Jakob (2025). *Champions League: Baskets Bonn verlieren Nervenschlacht gegen Reggiana*. Accessed: 14.01.2025. URL: <https://www1.wdr.de/sport/basketball/champions-league-baskets-bonn-spielbericht-100.html>.
- Höchli, Stefan (1981). *Zur Geschichte der Interpunktion im Deutschen. Eine kritische Darstellung der Lehrschriften von der zweiten Hälfte des 15. Jahrhunderts bis zum Ende des 18. Jahrhunderts*. Berlin/New York: De Gruyter.
- Institut für Deutsche Sprache (2018). *Korpusbasierte Zeichenhäufigkeitslisten. DeReChar v-uni-XXX-2018-02-28-1.0. Benutzerdokumentation*. Accessed: 03.01.2025. URL: <https://www.ids-mannheim.de/fileadmin/kl/derewo/derechar-v-uni-XXX-2018-02-28-1.0.pdf>.
- Jones, Bernard (1995). "Exploring the Variety and Use of Punctuation". In: *Proceedings of the 17th annual conference of the Cognitive Science Society*. Ed. by Johanna D. Moore and Jill Fain Lehman. Mahwah/Howe: Erlbaum, pp. 619–624.
- Klein, Wolf Peter (1998). "Über Schriftpartikel, oder: Warum man manchmal aus einer Mücke einen Elefanten machen darf". In: *Particulae particularum*. Ed. by Theo Harden and Elke Hentschel. Tübingen: Stauffenburg, pp. 177–186.
- Klockow, Reinhard (1978). "Anführungszeichen, Norm und Abweichung". In: *Linguistische Berichte* 57.1, pp. 14–24.
- Kultusministerkonferenz (2012). *Bildungsstandards im Fach Deutsch für die Allgemeine Hochschulreife*. Accessed: 14.01.2025. URL: https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Deutsch-Abi.pdf.
- Langlotz, Miriam (2017). "'Der zweite Nebensatz lässt sich besser als einzelner Hauptsatz lesen'—Interpunktionskorrekturen in Schülertexten". In: *Interpunktion im Spannungsfeld zwischen Norm und stilistischer Freiheit. Sprachwissenschaftliche, sprachdidaktische und literaturwissenschaftliche Perspektiven*. Ed. by Kristin George et al. Frankfurt am Main: Lang, pp. 125–156.
- Lauchman, Richard (2010). *Punctuation at work: simple principles for achieving clarity and good style*. Saranac Lake: American Management Association.
- Levin, Magnus and Jenny Ström Herold (2021). "On brackets in translation (or how to elaborate in brackets)". In: *Bergen Linguistic Studies* 11.1, pp. 121–144.
- Ludwig, Otto and Eckehart Merchert (1987). "Fritz Rahn und der Besin- nungsaufsatz. Zur Herkunft und Bedeutung einer Aufsatzform". In: *Praxis Deutsch* 84.1, pp. 10–16.
- Masalon, Kevin Christopher (2014). *Die deutsche Zeichensetzung gestern, heute—und morgen (?): eine korpusbasierte, diachrone Untersuchung der Interpunktion als Teil schriftsprachlichen Wandels im Spannungsfeld von Textpragmatik, System und Norm unter besonderer Berücksichtigung des Kommas*. Ac-

- cessed: 18.11.2024. URL: <https://nbn-resolving.org/urn:nbn:de:hbz:464-20140506-100648-7>.
- Meier-Vieracker, Simon (2021). *14 Monate DLF-Corona-Liveblog. Ein korpus-linguistisches Profil der COVID19-Pandemie*. Accessed: 14.01.2025.
- Meletis, Dimitrios (2020a). *The Nature of Writing. A Theory of Grapholinguistics*. Vol. 3. Grapholinguistics and Its Applications. Brest: Fluxus Editions.
- (2020b). “Types of allography”. In: *Open Linguistics* 6.1, pp. 149–266.
- Meletis, Dimitrios and Jonas Romstadt (n.d.). “The world’s punctuation systems: Comparing forms, functions, and variation”. In: *Comparative Punctuation Worldwide*. Ed. by Anna Saller and Paul Rössler. Accepted. Berlin/Boston: De Gruyter.
- Mentrup, Wolfgang (1983). *Zur Zeichensetzung im Deutschen—Die Regeln und ihre Reform. Oder: Müssen Duden-Regeln so sein, wie sie sind?* Tübingen: Narr.
- Niggemeier, Stefan (2014). *Problemformat Nachrichtenticker. Live dabei, wenn nichts passiert*. Accessed: 27.09.2024. URL: <https://www.faz.net/aktuell/feuilleton/medien/stefan-niggemeier-ueber-live-ticker-12859289.html>.
- Nunberg, Geoffrey (1990). *The Linguistics of Punctuation*. Stanford: CLS.
- Parkes, Malcolm B. (1993). *Pause and Effect. An Introduction to the History of Punctuation in the West*. Berkeley/Los Angeles: University of California Press.
- Primus, Beatrice (1997). “Satzbegriffe und Interpunktion”. In: *Zur Neuregelung der deutschen Orthographie. Begründung und Kritik*. Ed. by Gerhard Augst et al. Tübingen: Niemeyer, pp. 463–487.
- Ransmayr, Jutta (2022). “Schulische Rechtschreibleistungen unter Beobachtung—Schlaglichter aus der Untersuchung eines österreichischen Deutschmatura-Korpus”. In: *Sprachreport* 38.2, pp. 42–47.
- Rat für deutsche Rechtschreibung (2024). *Amtliches Regelwerk der deutschen Rechtschreibung. Regeln und Wörterverzeichnis*. Accessed: 04.07.2024. URL: https://www.rechtschreibrat.com/DOX/RfdR_Amtliches-Regelwerk_2024.pdf.
- Reh, Sabine, Michael Kämper-van den Boogaart, and Joachim Scholz (2017). “Eine lange Geschichte: Der deutsche Abituraufsatz als ‘Gesamtbildung der Examinanden’. Prüfungspraxis und Lehrerkommentare von Abituraufsätzen in den 1950er Jahren”. In: *Zeitschrift für Pädagogik* 63.3, pp. 280–298.
- Reißig, Tilo (2015). *Typographie und Grammatik. Untersuchungen zum Verhältnis von Syntax und Raum*. Tübingen: Stauffenburg.
- Rödel, Michael and Björn Rothstein (2015). “Die Kategorie Tempus, der Begriff der Funktion und ihre Didaktik”. In: *Was tun mit dem Verb? Über die Möglichkeit und Notwendigkeit einer didaktischen Neuerschließung des Verbs*. Ed. by Birgit Mesch and Björn Rothstein. Berlin/Boston: De Gruyter, pp. 219–250.

- Romstadt, Jonas (n.d.). *Variation im Gebrauch von Interpunktionszeichen. Untersuchungen in Abiturklausuren zwischen 1948 und 2018*. In preparation.
- Romstadt, Jonas, Theresa Strombach, and Kristian Berg (2024). "Graph-Var—Ein Korpus für graphematische Variation (und mehr)". In: *Orthographie in Wissenschaft und Gesellschaft. Schriftsystem — Norm — Schreibgebrauch*. Ed. by Sabine Krome et al. Berlin/New York: De Gruyter, pp. 425–435.
- Rössler, Paul, Peter Besl, and Anna Saller, eds. (2022). *Vergleichende Interpunktion—Comparative Punctuation*. Berlin/Boston: De Gruyter.
- Schreiber, Niklas Heiner (2020). *Die Syntax des Semikolons. Von links ein Punkt—nach rechts ein Komma*. Berlin: Springer.
- Schüler, Lisa (2023). "Interpunktion im Medienvergleich: Untersuchungen zu handschriftlichen, tastaturgeschriebenen und am Computer diktierten Texten". In: *Lesen- und Schreibenlernen im Spannungsfeld zwischen Wissen und Können*. Ed. by Mangel Bangel and Iris Rautenberg. Baltmannsweiler: Schneider, pp. 81–114.
- Sebba, Mark (2007). *Spelling and society: the culture and politics of orthography around the world*. Cambridge: Cambridge University Press.
- Sieber, Peter (1998). *Parlando in Texten. Zur Veränderung kommunikativer Grundmuster in der Schriftlichkeit*. Tübingen: Niemeyer.
- Ström Herold, Jenny and Magnus Levin (2024). "From Dashes to Dashes?—A Corpus Study of Dashes in English, German and Swedish". In: *Contrastive Corpus Linguistics: Patterns in Lexicogrammar and Discourse*. Ed. by Anna Cermátová et al. London: Bloomsbury, pp. 260–284.
- Zifonun, Gisela, Ludger Hoffmann, and Bruno Strecker (1997). *Grammatik der deutschen Sprache*. Berlin/New York: De Gruyter.
- Zimmermann, Harald (1968). *Zur Leistung der Satzzeichen. Eine Studie über die Funktion der Zeichensetzung im Deutschen, untersucht am Beispiel der Gegenwartssprache*. Accessed: 18.11.2024. URL: <https://zimmermann.infowiss.net/pdf/1969a.pdf>.

Un court manuel du minoen proto-gréco-anatolien

Alexandre Solcà & Francesco Perono Cacciafoco

Résumé. Courte communication suite à notre présentation pour le colloque /gra-fematik/ 2024.

Le texte suivant est le résultat synthétique des recherches effectuées entre 2021 et 2025. La rédaction de cette étude est entièrement la mienne (A.S.) Nous sommes infiniment reconnaissants au professeur Perono de son soutien et à ma tendre épouse et mes parents pour leurs conseils, profonds échanges et confiance dans nos entreprises.

Cette année (2026), nous avons l'intention de présenter un court manuel du minoen proto-gréco-anatolien qui pourra éclairer plusieurs inscriptions et aspects toponymiques de la Crète actuelle et aussi de l'Âge du Bronze. Avant d'avancer dans des considérations plus détaillées, il nous semble nécessaire de présenter brièvement trois étapes essentielles dans la discussion du minoen : il existe, à notre avis, trois variantes principales de ce langage.

1. le minoen observé sur les sceaux minoens et les inscriptions hiéroglyphiques comme le disque de Phaïstos et dont la datation est assez haute dans la chronologie crétoise de l'Âge du Bronze ;
2. le minoen transcrit dans un système d'écriture phonétique mono- et plurisyllabique (appelé Linéaire A) en raison de sa proximité graphique avec le mycénien (appelé encore Linéaire B). La principale difficulté de cette étape du minoen est que la nature de ses inscriptions (tablettes, vases, sceaux) n'est pas toujours de la même source et donc leur interprétation est complexe et nécessite beaucoup de prudence. Il manque de plus un véritable lexique des mots en linéaire A

Alexandre Solcà  0009-0004-5797-0398
Chercheur indépendant Moscou-Lausanne
E-mail : aleksandrsolka@yandex.com

Francesco Perono Cacciafoco  0000-0002-0977-063X
Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, CN
E-mail : Francesco.Perono-Cacciafoco@liverpool.ac.uk

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN : 2681-8566, e-ISSN : 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 159-162. <https://doi.org/10.36824/2024-graf-solc>
ISBN : 978-2-487055-10-0, e-ISBN : 978-2-487055-11-7

- et leurs formes parallèles en anatolien et proto-grec, et voilà ce sur quoi nous travaillons actuellement ;
3. le minoen que l'on observe dans certaines tablettes en linéaire B, en égyptien et dans d'autres sources méditerranéennes et qui nous indique une étape de langue différente.

En raison de ces critères, il nous semble adéquat de parler non de minoen mais de langues égéo-anatoliennes sur terres crétoises.

Ainsi, il est plus aisé et précis d'analyser ces inscriptions sous la forme de langues d'une région ayant grandement développé son influence, notamment en raison du commerce et de la navigation et ayant perfectionné au cours des âges ses caractéristiques linguistiques propres.

Pour mieux comprendre ce qui fait la spécificité des langues égéo-anatoliennes en terres crétoises, il faut se représenter des princes et princesses crétois, certains originaires d'Anatolie occidentale, d'autres des terres égéennes, qui, unifiés sous un royaume dit minoen, bien que le terme devrait être corrigé à notre avis, organisaient l'ensemble de la Crète sous une large confédération de protecteurs et « vassaux ».

Pour les points concernant strictement les écritures crétoises, il nous paraît important d'utiliser trois catégories lors de la discussion du matériel archéologique et épigraphique en provenance de Crète ou d'îles égéennes en relation avec le monde minoen-anatolien.

Nous proposons donc ces trois abréviations :

- HMA (Hiéroglyphique Mino-Anatolien),
- EACPS (Égéo-anatolien crétois pluri-syllabique),
- CPE (Crétois post-égéen).

À noter qu'en anglais, celles-ci seront transcrites MAH, PSCEA, PEC.

Ainsi, la catégorisation de chaque inscription ou de chaque matériel portant une inscription s'en trouvera clarifiée et plus précise lorsque l'on procède à une analyse en détail de son contenu lexicologique et morphologique.

Si la majorité des publications en lien avec le minoen adoptait ce système, cela augmenterait les chances de déchiffrements précis et d'une meilleure compréhension du corpus égéo-anatolien en général, y compris en ce qui concerne les autres systèmes d'écriture de cette région, comme, par exemple, le dit chyro-minoen.

Passons maintenant en revue certains termes observés dans les lexiques HMA et EACPS.

Dans la majorité des tablettes transcrites en linéaire A (EACPS), on observe la présence de noms en lien avec la réalité géographique et culturelle de Crète (e.g., A-DI-KI-TE, le fameux mont Dikté), des mots en lien avec la généalogie mythologique (A-TA-I, cf. atta en hittite et

oreu en russe : papa) mais aussi des mots contenant des formes quasiment équivalentes au proto-grec (KU-RU-KU, cf. grec κρόκος).

Il devient alors bien plus évident comment approcher les subtilités lexicales du corpus égéo-anatolien crétois. Et aussi d'observer les phases de développement de ces différents langages ou plus précisément les phases dialectales d'un créto-anatolien.

Nous voudrions dans un deuxième volet exposer en détail la liste complète des mots que nous avons pu analyser, mais cela sera inclus dans notre futur manuel du minoen proto-gréco-anatolien.

Comme nous l'avions présenté en ligne, le crétois minoen possède plusieurs formes lexicales mais aussi graphiques et c'est l'étude approfondie de leurs parallèles et différences qui produit un résultat clair et exact de ces combinaisons phonographiques.

Je remercie le comité de la conférence *Grapholinguistics in the 21st Century* de son intérêt et de la compréhension de D^r Yannis Haralambous concernant notre dossier remis avec du retard.

L'article suivant comportera les images des tablettes et le lexique crétois égéo-anatolien dont nous avons exposé les principes.

Nous espérons pouvoir discuter prochainement avec d'autres spécialistes de ces questions et remercions nos lecteurs et lectrices de leur intérêt. Pour toute autre information ou demande de précision, vous êtes cordialement les bienvenus pour nous contacter par email.

Ressources

1. Introduction au Linéaire A par Alexandre Solcà, secrétaire général de la Societas Anatolica et spécialiste de linguistique antique et moderne <https://www.youtube.com/watch?v=dzYWn3MiSQo>
2. Les sceaux minoens et les écritures hiéroglyphiques de Crète : première partie, par Alexandre Solcà <https://www.youtube.com/watch?v=ZWmPahZ7b-o>
3. Les sceaux minoens : écriture syllabo-hiéroglyphique du monde égéo-anatolien (seconde partie du précédent exposé d'Alexandre Solcà) <https://youtu.be/i2I7SerkrYc?si=pJLR4av0gNB3CizK>

Références

- “A versatile Mycenaean scribe” (2021). National Archaeological Museum, https://www.namuseum.gr/en/monthly_artefact/a-versatile-mycenaean-scribe/.
- BEST, J.G.P. (1982–1983). “The Zakro Pithos Inscription, Again”. In : *Talanta* 14–15, p. 9–16.

- DOSUNA, Julián Méndez, Thomas G. PALAIMA et Carlos Varias GARCÍA, éd. (2022). *Studies in Mycenaean Texts, Language and Culture in Honor of José Luis Melena Jiménez*. T. 94. Hellenic Studies. Harvard University Press.
- LANDENIUS ENEGREN, H. (2008). *The People of Knossos : Prosopographical Studies in the Knossos Linear B Archives*. Uppsala : Uppsala Universitet.
- ΠΛΑΤΩΝ, N. [Platon, N.] et W.C. BRICE (1975). *Ενεπίγραφοι Πινακίδες καὶ Πῖθοι Γραμμικοῦ Συστήματος Α ἐκ Ζάκρου [Inscribed Tablets and Pithos of Linear A System from Zakro]*. T. 85. Βιβλιοθήκη τῆς ἐν Ἀθήναις Ἀρχαιολογικῆς Ἑταιρείας [Library of the Archaeological Society at Athens]. Athens : Ἀρχαιολογικὴ Ἑταιρεία Ἀθηνῶν [The Archaeological Society at Athens].

Inscribing Multilingualism

Lifemaps and “Trans-scripting” Among Moroccan Amazigh Youth

Adeli Block

Abstract. In this paper, I examine script and writing practices after Tamazight officialization and standardization in the broader linguistic landscape of Morocco. Drawing on over 150 lifemaps (García-Sánchez 2014), whereby students in Tamazight classes in public elementary schools (grades 1 through 6) draw the places they frequent and label them, I consider script contact (contact between scripts) and the emergent ways students in southeast Morocco are expressing themselves multimodally with their entire semiotic repertoire. Similar to translanguaging (Otheguy et al. 2015) that recognizes peoples’ full linguistic repertoire without reifying named (colonial) languages, the lifemaps demonstrate the ways in which young students are, what I’m calling, *trans-scripting* or *transgraphing*. For example, elementary school students utilize several scripts (Neo-Tifinagh, Latin, and Arabic), write Arabic words with Tifinagh script, write Tamazight words with Arabic script, and muddle script directionality. I aim to take the aesthetics of scripts and children’s drawings seriously as an instantiation of linguistic justice (Baker-Bell 2020) that recognizes multiple ways of being “correct” as students are being interpellated as multiscrptal and multilingual citizens in the Moroccan social imaginary. Visually, I analyze the different aesthetics of frequented communal places. Indubitably, lifemaps are a powerful act of self-representation that reveals the ways in which young minds interpret spaces and places. Crucially, I contend that these lifemaps, with their distinctive aesthetics, function as an “other-archive” (El Guabli 2023) that documents a history enmeshed in loss and silence against the backdrop of layers of colonialism.

In January 2023, I held hands with a seven-year-old girl in southeast Morocco, whom I will call Amira. We were about to enter the weekly *souq* (market), and we saw her French teacher nearby, and we waved. I said to her, “*is trẓrit aslmad nnm?*” (Did you see your teacher?). I said “teacher” using the Tamazight word. She said, “No! That is not *aslmad*,” emphatically. “That’s *monsieur* or *ustad*.” For her, *aslmad* (the Tamazight word for “teacher”) was reserved only for the Tamazight teacher, not just any teacher. A few weeks later, I was observing Amira and her mom, Fatima, work through some homework, which her mom was very invested in.

Adeli Block  0009-0007-2849-0186

Department of Anthropology, University of Michigan, Ann Arbor, MI, US
E-mail: avblock@umich.edu

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 163–179. <https://doi.org/10.36824/2024-graf-bloc>
ISBN: 978-2-487055-10-0, e-ISBN: 978-2-487055-11-7

Amira had drawn a rainbow and some other landscape features, and she needed to come up with a title. I suggested, “How about you name it *tislit n unẓar*?” referring to “rainbow” in Tamazight (lit. bride of the rain). Her mom quickly interjected, “No, this is for Arabic; it has to be in Arabic.” I told her daughter, “Okay, so write *tislit n unẓar* in Arabic letters.” Her mom said, “No.” They settled on *nuḥaaḥḥidh ‘ala al-bi’a* in Modern Standard Arabic, “let’s preserve the environment.” These two moments exemplify some of the ways in which children are socialized into conceptualizing language boundaries and script boundaries, both reinforced by school and at home, through their parents. Arabic script is for Arabic, Latin is for French, and Tifinagh script is for Tamazight. In other words, there is a conflation between language and referential content, as we see with *aslmad* and *monsieur*, and there is a conflation between script and language where Arabic script is meant only for Arabic words rather than a medium to write any language. My paper today looks specifically at writing, script, and the visual practices of children in southeast Morocco, and the different ways kids are reproducing this script essentialism and also subverting it.

In 2001, through a royal decree, King Mohamed VI created the Royal Institute for Amazigh Culture (also called IRCAM) to promote Amazigh languages and begin to codify and standardize them. In 2002, some schools first began teaching Tamazight, after decades of marginalization and erasure. After generations of activism and resistance, the Amazigh Culture Movement achieved a milestone in 2011 when Tamazight became an official language alongside Arabic in the wake of the “Arab Spring.” Notably, the King, in coordination with IRCAM, selected Tifinagh, not Latin nor Arabic, as the official script for Tamazight, which is an ancient script that they modernized and revived. This government choice scripted distinction and distinctiveness because Latin indexed French, and Arabic indexed Arabism/Islamism. This schema results in an essentialized and conflated relationship between script and language.

For example, on a prominent bank’s ATM screen (Figure 1) asking, “choose your language,” Tifinagh, **the script**, is listed as the language rather than Tamazight, the actual name of the language. School then becomes a key site through which students are taught to separate these codes and scripts while also being interpellated as multilingual and multiscrptal citizens.

In Figure 2, we see the days of the week in the three separate codes (Arabic, Tamazight, and French) with their respective essentialized script (Arabic, Tifinagh, and Latin). I found this hanging in one of the classrooms at another nearby school that teaches Tamazight. While Tamazight and Arabic are the official languages, French continues to be taught in schools and for grades 3-6, students are taught six hours of French plus math and science (also taught in French) compared to the three hours allotted for Tamazight. Currently, Tamazight is only be-

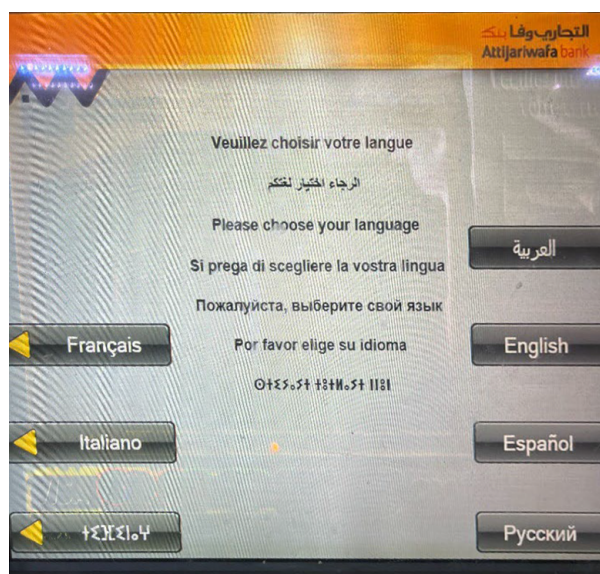


FIG. 1. Bank Screen for Attijariwafa Bank

ing taught in some elementary schools, with middle schools to soon roll out. Teacher shortage continues to be a crucial problem in universalizing Tamazight education across the country. As a scholar of language and script politics, Tamazight classes in elementary school classrooms in the southeast became a prominent setting for my research.

In designing my dissertation fieldwork, I was inspired by linguistic anthropologist Inma Garcia-Sanchez's use of lifemaps with Moroccan immigrant children in Spain (2014). While her method combined drawings and text as acts of self-representation, my approach differs by integrating lifemaps more systematically with children in Al-Jadida elementary school, rather than just focal children (see Figure 3).

In collaboration and consultation with the Tamazight teacher Achraf, we decided to include the lifemap activity as a structured classroom activity that all children in the school would participate in. For grades 1 through 4, the lifemaps were done during the *nasbat tarfibi* (the fun activity) as part of the Tamazight class. For grades 5 and 6, participants in the Tamazight club also drew lifemaps. When Achraf introduced the activity to the students, their faces lit up. They were so excited to see that they would have the opportunity to draw, and on actual drawing paper, as part of class. Parents and teachers alike were surprised that I spent money to buy the "special drawing paper" but appreciated it by saying, "bravo 'alik!" I told them that it was not my personal money but money acquired through research funding for these exact kinds of expenses. Ad-



FIG. 2. Days of the week in Arabic, Tamazight, and French

ditionally, the students were central to my research, and so I was thrilled to give them a “fun” activity, which I also knew would be fruitful in displaying heterogeneous writing and drawing practices. News travelled fast throughout the school, and classes already knew what was coming when it was their “turn” to do the lifemaps. Achraf would often take a student as an example to demonstrate to the students what we were looking for: “First, you wake up, and then where do you go?” “School!” “Great, and then after school, where do you go?” “Draw the places that you go to many times during the week, and then write down the names of the places. You can write the names however you want, or however you say them.” Often, he would draw a house on the board with an arrow to the next place to give them an idea of the task. We would remind students many times, “Don’t forget to write the names of the places! In any language!” Some students, when going over the directions, were more pre-occupied with it being “accurate,” especially in levels 3 and up. “How do I say classroom in Tamazight?” A couple of students said, “Can I write in the American language?” Achraf thought this was funny because they didn’t call it English. Achraf said, “You can write in Chinese,” signally that literally any code was permissible, even though they didn’t know these codes. Some students sought my approval: “Is this beautiful?” Some wanted to make it a competition for who could have the best



FIG. 3. An example of a lifemap

lifemap, given that students were accustomed to different competitions at school. I assured them that all maps are beautiful and that whatever they turned in would be great.

In this talk, I aim to take children's drawings and their writing practices "seriously" as both a powerful act of self-representation and as what Amazigh studies scholar, Brahim El Guabli, calls an "other-archive." Other-archives are:

texts, artifacts, alphabets [...] where stories of the excluded, the silenced, and the forgotten live in a ghostly state [...] other-archives are part of unfolding memory and history, existing to bridge the gap between a past that is not yet finished and a society still impacted by the consequences of the unfinished past". (El Guabli 2023: 1)

I argue that children's lifemaps should be treated as an "other-archive" in the afterlives of language policy change, not merely because it documents a history enmeshed in silence in loss through the reclamation of Tifinagh but also because the act of writing in Tamazight, in any script, and depictions of lifeworlds at the margins of Moroccan society constitute an act of resistance. As a new generation that is able to write in

three different languages and three different scripts emerges, these children are hailed by the state as multilingual and multiscriptal citizens, while also subverting hegemonic institutionalized practices that privilege Arabic (and French). Lifemaps are also central to my ethnographic study of script politics because they engender the three rights or principles that shape Amazigh identification. These are: *akal* (land), *afgan* (person), and *awal* (speech/language). The Amazigh Cultural Movement invented or reinvented these principles as integral to Amazighness. Thus, these lifemaps uniquely depict the nexus of land, people, and language as children draw the places they frequent and label them.

While some scholars and instructors would consider these semiotic practices as “wrong,” “incorrect,” or “truncated” (Blommaert 2010), I aim to valorize them as acts of resistance and self-expression against the backdrop of layers of colonialism and linguistic oppression. Similar to how scholars in linguistics have theorized translanguaging as a framework that rejects (colonial) named languages that often correspond to nation-states (Otheguy 2015 et al.), I understand these linguistic and script practices to be “trans-scripting.” Thus, while the state, through schools, is trying to socialize students into a script or graphic regime where Arabic=Arabic, Latin=French, and Tifinagh=Tamazight, these students demonstrate the de-essentialized relationship between script and language. In other words, while scripts are essentialized (rheumatized) by the state, and reified through elementary schools and the standardized curriculum, kids are subverting this by writing words in whatever script they want, as we will see in these lifemaps.

Here is an example of a lifemap (Figures 4 and 5) where we can observe script contact and language contact from a first grader. “Police” is written in French in Latin script, “*tinml*” (“school” in Tamazight) is written in Tifinagh script, and she has her name and “field” written in Arabic.

These lifemaps are a point or stoppage, at a particular time on their script socialization trajectory, through which we can understand the semiotic ideological assemblages in which students are enmeshed. We can assess how students are thinking about the relationship between script and code, different varieties (Moroccan Standard Tamazight, Standard Arabic, Moroccan Arabic, local Tamazight, and French), and their lifeworlds. Elementary school is an ideal site to do these lifemaps because it is this very space where students are acquiring three scripts and 3+ codes simultaneously.

After coding 173 of these maps thus far, here are some of the significant findings. The main kinds of places that children drew are school, home, the store, the mosque, a friend/family house, and soccer fields.

One salient feature is the caricature-like nature of many of the children’s drawings, especially when houses are drawn. Many students, in particular, drew box-shaped homes with triangular roofs, undoubtedly



FIG. 4. The frontside of a first grader's lifemap



FIG. 5. The backside of a first grader's lifemap

images of homes that they have encountered in cartoons, movies, and other media, given that no homes locally look like that.

However, some students took a more emic or locally constructed approach and aimed to represent southeastern architecture where structures are adorned with tower corners as depicted in Figure 6. This is significant for depicting what their actual local architecture looks like rather than reproducing images that are circulated in the media.

Some students' images of the mosque were particularly detailed, indexing degrees of piety or the importance of Islam to their lifeworld constructions. 64% of students drew a mosque on their lifemap.

For example, several students wrote *Allabu 'akbar* (in Arabic script), "Allah is the greatest," which is how the call to prayer begins and ends. Moreover, certain students wrote *makaan lwaḍu'* (place for ablution), indexing their high degree of familiarity with space and signage in mosques.

For some, kids drew *ḥanut/taḥanut/maḥal/dukkān* (convenience stores) because running to the convenience store on behalf of mom is an integral part of many children's roles in the family.

Many children, especially boys, drew a soccer field, even those who don't play soccer. Nearly 50% of the boys drew a field compared to 15% of girls. I surmise that this is mainly for two reasons. Firstly, there is an impressive turf complex adjacent to the school for club teams; thus, the field is a salient place in their everyday geography, whether or not they utilize it. Second, soccer is an ever-growing sport in Morocco for all genders, with both the men and the women having significant World Cup and Olympic runs. Nonetheless, soccer continues to be a male-dominated sport.

Many students also drew gardens. In the center or downtown part of Tinjedad, there is a garden that has swings, a playground spinner, many benches, popcorn/chip/ice cream vendors, and so forth. Many families spend their evenings in this park area as a recreational activity, and it's also a public space where women can convene and chat. Trees were also present in many of the children's maps because Tinjedad is a pre-Saharan oasis town, and so there are many fields filled with palm trees on either side of the main road that cuts through the center of Tinjedad.

In these lifemaps, we see students who have drawn very specific places that are tied to their "asl" or origin, which is a key principle of social organization in southeast Morocco.

Depicted here are different fortified villages that connect to their family's lineage, demonstrating the spatial networks that children are embedded in.

I will now present a map from each grade level so I can discuss some of the key patterns that are featured across these maps. First, we have Hamza, who is a first grader.



FIG. 6. Examples of students who incorporated southeastern architecture in their lifemaps

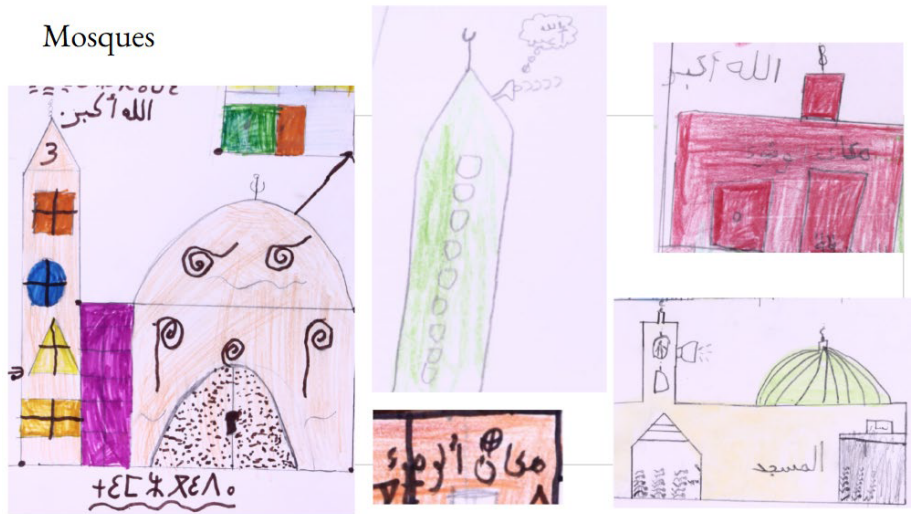


FIG. 7. Examples of mosques drawn on students' lifemaps

Fortified Villages



FIG. 8. Fortified villages and other cities drawn on children's lifemaps

His name is written in Tifinagh. He starts his day off at *al-manzil* (house) in Arabic but written in Tifinagh script. Then he goes to school, *tinml*, the Tamazight word for “school.” Then he goes to a garden, *turtit*, also in Tamazight, then he goes to a mosque, *timzgida* also in Tamazight, then he goes to the library, *al-maktaba* written in Arabic but with Tifinagh letters, and then he goes to *taddart n Yousef*, (Youssef’s house) written in Tamazight. Clearly, we have a lot of mixing here. Next, we have a second grader.

He has *hadiqa* (garden) in Standard Arabic, *tinml* in Tamazight (and Tifinagh script), *taddarat* meant to be *taddart* in Tamazight, and *nk ligh ghor baba* (I’m at dad’s), a sentence in Tamazight written in Arabic script.

For this third grader (Figure 11) he starts with *tadart*, the Tamazight word for “home,” then he goes to *tinml* the Tamazight word for “school,” then he goes to *mal’ab*, the Standard Arabic word for “field,” then the *masjid* or the “mosque” in Arabic, then the *jarda* (garden) which is the Darija or Moroccan Arabic term (borrowed from the French *jardin*), and then *piscine* (pool) written in Arabic letters.

For this fourth grader, we see only tifinagh script, but as you may have surmised, she does not use only tamazight words.

For “mosque,” she uses the Arabic *ljaam* and for “library,” she uses the Arabic *lmaktaba* even though all her script choices are in Tifinagh. For grade 6, we have Mariam.

“House” and “school” are both in Tamazight and written in Tifinagh, “park” and her “remediation school” are in Arabic, and the café is in

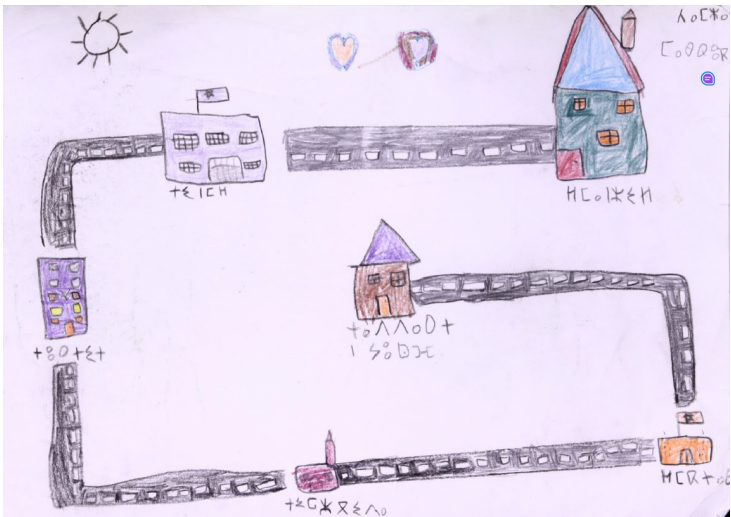


FIG. 9. A lifemap by a first-grade boy



FIG. 10. A lifemap from a second-grade boy

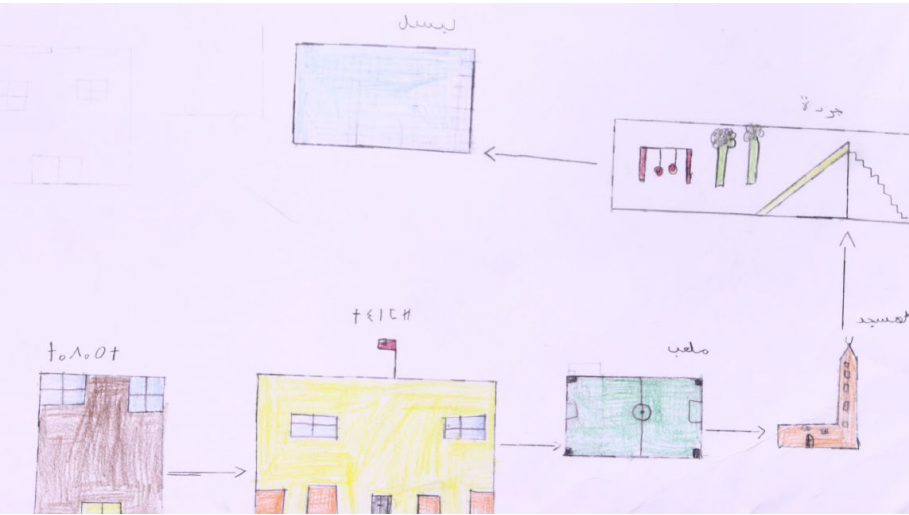


FIG. 11. Lifemap by a third-grade boy



FIG. 12. Lifemap by a fourth-grade girl



FIG. 13. A lifemap from a sixth-grade girl

Latin/French, which is called café Alaska, the new hot place to be in town.

This activity also allowed those who are most marginalized socially in the classroom to be able to participate in a classroom activity without judgment or fear of punishment. On Figure 14, I have compiled a few examples that illustrate this.

For example, we see a tent drawn by a second-grader who lives in the pastoral nomadic community near the school. Notably, he did not draw a caricature-looking house and instead drew the quarters in which he lives. Additionally, several students at the school have severe cognitive disabilities, and they are “integrated” with those who do not have cognitive disabilities as part of a change in education policy. While they are integrated nominally, they are often relegated to the back and occupy themselves, rarely being consciously included in the standardized curriculum. I have included several drawings here from students with cognitive disabilities across different grade levels. While the teachers often characterize their work as *walu* or “nothing,” I wanted to include their drawings because they do, in fact, matter. We see from Houda, a sixth grader (upper left photo in Figure 14), that she writes her name in Ara-



FIG. 14



FIG. 15. Examples of students muddling script directionality

bic, and also writes various letters in Tifinagh. While this does not mean anything referentially, we see that she is still engaging in a powerful act of “trans-scripting” freely and drawing on her entire semiotic repertoire.

When I’ve observed script mixing, there are three main kinds that emerge. The first is muddying script directionality.

In Figure 15, there are examples where the Arabic text is going left to right instead of right to left, and the Tamazight is going right to left instead of left to right. While this happens more at the younger levels, there were students in the upper levels who also showed this practice.

The second type of mixing is Arabic words in Tifinagh script.

Arabic Words in Tifinagh Script



FIG. 16. Examples of children writing Arabic words in Tifinagh script

The primary examples that occurred are the ones shown here *lmanzil* (house), *tirran*, a Moroccan Darija word coming from the French *terrain*, *ljaam*, the Arabic word for “mosque,” *hadiqa* the Arabic word for “garden,” and *madrassa*, the Arabic word for “school.” When Achraf asked Hamza why he wrote *lmanzil*, an Arabic word in Tifinagh, Hamza said, “I wanted to.”

The third type of mixing is Tamazight words in Arabic script.

The overwhelming Tamazight term written in Arabic script was *taddart*, meaning “house.” Other examples included here are *amdakkl* (friend), *tahanut* (store) and *tarbat* (girl). When Achraf asked Abdullah why he didn’t write *taddart* using Tifinagh letters, he said *ur tn sinngb* “I don’t know them.” Noticeably absent is French. Only one student labeled

Tamazight Words in Arabic Script



FIG. 17. Examples of children writing Tamazight words in Arabic script

a place in French, school, and he did so alongside *madrassa* and *tinml*, the Arabic and Tamazight words for school, as seen in Figure 18.

All 3 Codes (all 3 scripts)

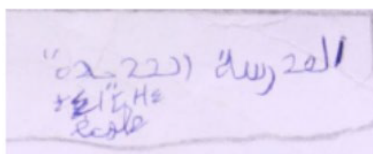


FIG. 18. "School" written in three codes and three scripts

Students are educated formally in French, but when given the option to write "freely," French was not a code that agentively chose.

As I have demonstrated, lifemaps are not clear-cut, and neither are people's lives. As a powerful act of self-representation and resistance against the backdrop of layered colonialisms, we must take lifemaps seriously as a form of social action. With more than 64% of student lifemaps being characterized as showing significant mixing, we must de-essentialize our understanding of codes, scripts, and languages. As the late linguist Jon Henner said, **"Don't let anyone tell you your language is wrong. Your languaging is the story of your life."** While scripts can

script distinction or distinctiveness, we cannot use them as a means to police, censure, or constrain people when we have a system that produces multiscritpal citizens. Just as translanguaging is a radical act in pursuit of linguistic justice (Baker-Bell 2020) that doesn't adhere to named language boundaries, trans-scripting or transgraphing as an anti-essentialist framework must also be included in our efforts toward language as liberation.

References

- Baker-Bell, April (2020). *Linguistic justice: Black language, literacy, identity, and pedagogy*. Routledge.
- Blommaert, J. (2010). *The sociolinguistics of globalization*. Cambridge University Press.
- El Guabli, Brahim (2023). *Moroccan Other-Archives: History and Citizenship after State Violence*. Fordham Univ Press.
- García-Sánchez, Inmaculada Ma (2014). *Language and Muslim immigrant childhoods: The politics of belonging*. John Wiley & Sons.
- Otheguy, Ricardo, Ofelia García, and Wallis Reid (2015). "Clarifying translanguaging and deconstructing named languages: A perspective from linguistics". In: *Applied linguistics review* 6.3, pp. 281–307.

Creating an Orthography for an Endangered Language: The Graphisation of Cypriot Arabic

Spyros Armostis & Marilena Karyolemou

Abstract. Graphisation, i.e., the development of a signary and associated orthographic conventions for a previously unwritten language variety, constitutes an important component of language corpus planning, particularly in the context of endangered-language revitalisation. This paper presents the graphisation of Cypriot Arabic, a severely endangered Arabic variety spoken by Maronite Cypriots of Kormakitis. Following its recognition as a minority language of Cyprus in 2008 and the establishment of revitalisation initiatives, the need arose for a codified writing system to support language documentation, teaching, and community use. A proposed signary and orthography were developed on the basis of linguistic analysis and subsequently evaluated through a study involving native speakers. The study investigated preferences regarding script choice, orthographic depth, and the representation of selected morphophonological phenomena, as well as attitudes towards the proposed system in terms of acceptability, usability, and learnability. Participants generally considered the proposal acceptable and logical. However, they did not uniformly favour either phonetic transparency or morphophonological depth. Instead, preferences varied according to the linguistic phenomenon under investigation, with surface-oriented spellings preferred for lexical processes and deeper spellings for post-lexical ones. The findings suggest that successful graphisation should combine linguistic analysis with speaker feedback and that orthographic depth may vary across phenomena rather than characterising a writing system as a whole.

1. Introduction

Linguists are often called upon to propose a writing system for language varieties that have hitherto existed primarily as oral traditions with no established written norm. In such graphisation efforts, especially in cases of endangered varieties, multiple parameters must be taken into account in order for the implementation of a new system to have the potential to succeed. Beyond the linguistic soundness of the proposed or-

Spyros Armostis  0000-0003-2927-2096
University of Cyprus, E-mail: armosti.spyros@ucy.ac.cy

Marilena Karyolemou  0000-0003-0914-5794
University of Cyprus, E-mail: makar@ucy.ac.cy

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 181–207. <https://doi.org/10.36824/2024-graf-armo>
ISBN: 978-2-487055-10-0, e-ISBN: 978-2-487055-11-7

thography, issues of acceptability and learnability among native speakers of the variety become central. The present study reports the results of an investigation that tested a proposed orthography for Cypriot Arabic, a severely endangered language in the mediterranean island of Cyprus.

1.1. Cypriot Arabic

Cypriot Arabic (endonym: *sanna* ‘our language’, also referred to as *arapika* ‘Arabic’ within the community) is spoken by members of the Maronite community of Cyprus, one of the constitutionally recognised ethnoreligious groups of the Republic of Cyprus. In the scientific literature the variety is known as *Cypriot Arabic*, while the designation *Cypriot Maronite Arabic* has appeared to a certain degree in official contexts; the informal label *marōnitika* ‘Maronite language’ is also attested but imprecise (Armostis, 2025). Maronite Cypriots belong to the Maronite Church, an Eastern Catholic Church based in Lebanon with Saint Maron as its patron saint. Maronite populations arrived on the island in successive waves from the region of Greater Syria between approximately the 7th and 12th centuries, with medieval sources reporting the presence of numerous Maronite villages especially to the north of the island (Frangeskou and Hadjilyra, 2012, p. 8). Today only four Maronite villages remain, all located in the northern part of the island. Cypriot Arabic is spoken exclusively by Maronites originating from the Kormakitis village; Maronites from other villages neither speak the variety nor have any memories of the language ever having been used in their villages in the past.

From a typological and historical perspective, Cypriot Arabic belongs to the group of peripheral Arabic varieties, i.e. varieties that developed outside the core Arabic-speaking area and in sustained contact with non-Semitic languages (Borg, 1985; 1994). In this case, prolonged contact with Cypriot Greek has resulted in extensive lexicogrammatical influence (Armostis and Karyolemou, 2023). Following the turbulent events of 1974 and the displacement of the Kormakitis population, community networks were disrupted and intergenerational transmission virtually arrested. As a result, Cypriot Arabic is today spoken as a first language by fewer than one thousand speakers. Therefore, Cypriot Arabic is classified as severely endangered in UNESCO’s *Atlas of the World’s Languages in Danger* and is characterised by very low ethnolinguistic vitality (Karyolemou, 2019).

The official recognition of Cypriot Arabic as a minority language of the Republic of Cyprus in 2008, following joint efforts by academics and members of the Maronite community, constituted a crucial step in formal language planning, and more specifically in status planning: this recognition established Cypriot Arabic as a language in its own right rather than as a mere dialectal form of Arabic, thereby raising its insti-

tutional profile and enabling access to funding and coordinated protection measures (Armotis, 2025). In consultation with community representatives, a team of academics subsequently designed a structured programme for documentation and maintenance/revitalisation, while the government appointed an Experts Committee composed of academics and community members to advise on revitalisation strategy. These developments culminated in 2013 in the launch of the 'Research Programme for the Documentation, Preservation and Revitalisation of Cypriot Maronite Arabic' under the Cyprus Ministry of Education (the now-called the Ministry of Education, Sports and Youth), a programme in which the present authors participate (*ibid.*).

Given that Cypriot Arabic had for centuries functioned almost exclusively as an orally transmitted variety, with very limited written material and no comprehensive grammatical or lexicographic description, the programme initially prioritised systematic documentation as a foundation for corpus planning (Armotis, 2025; Armotis and Karyolemou, 2023). Documentation, however, was conceived not as an end in itself but as a prerequisite for maintenance and revitalisation. From the outset, the development of teaching and learning materials emerged as a central objective. This, in turn, made explicit the absence of a codified orthography for Cypriot Arabic and necessitated its practical development, i.e., its *graphisation*.

Graphisation is part of language corpus planning (Fishman, 1974; Haarmann, 2006; Liddicoat, 2005) and can be defined as the development of a signary and associated orthographic conventions for a given language variety (Fenyvesi and Bakrá-Nagy, 2025; Haarmann, 2006). The signary is 'the complete inventory of the basic signs of a given writing system' (Coulmas, 2003, p. 36), which includes not only signs corresponding to language units (distinctive features, phones, phonemes, syllables, moras, or morphemes/words), but also other symbols, such as typographic blank spaces, punctuation marks, and numerals. Signary development, as a component of graphisation, is called *alphabetisation* in the case of alphabetical writing systems (Haarmann, 2006); for graphisation to be complete, the creation of an associated orthography is required. Orthography is an optional component of a writing system and is understood as regulation of the system by means of prescriptive rules specifying how one may write correctly in the given writing system (Meletis and Dürscheid, 2022, p. 171–172).

1.2. Writing in Cypriot Arabic

In discussing the development of a signary and orthography for Cypriot Arabic, it is necessary to consider the literacy backgrounds of speakers and their familiarity with existing scripts and writing systems. Mem-

bers of the Cypriot-Arabic-speaking community have native competency not only in Cypriot Arabic, but also, in recent years, in Cypriot Greek and are literate in Standard Modern Greek; they therefore have full command of the Standard Modern Greek writing system and the Greek script in general. In addition, through formal education, many speakers have learnt English, French, and/or Italian as foreign languages and are consequently familiar with their writing systems and with the Roman script more broadly. By contrast, there is practically no literacy in Modern Standard Arabic within the community, with the exception of a couple of individuals. As a result, the Arabic script is largely unfamiliar and not functionally available for community-wide literacy practices. Syriac abjad is used in liturgical contexts, but it is not directly accessible to most Cypriot Arabic speakers, since liturgical texts are typically mediated through Greek transliteration rather than read in the original script.

On the basis of the above considerations, any graphisation effort would be more successful if it was based on the Greek and/or Roman scripts rather than on the Arabic one. This is supported by the limited writing practices in Cypriot Arabic that existed prior to the graphisation efforts of the above-mentioned research programme. For instance, Cypriot Arabic was rendered in writing with the use of the Greek script in the Frangiskou (2000) dictionary and in some other unpublished material. Even though the use of the Greek script (with the addition, for some users, of two letters from the Roman script, namely and <d>) was not conceived necessarily as a full alphabet proposal in those sources, it can be summarised as follows:

- (1) <Αα Ββ Γγ Δδ Εε Ζζ Ζζ Ηη Θθ Ιι Κκ Λλ Μμ Νν Ξξ Ξξ Οο Ππ Ππ/Ββ Ρρ Σσς Σσς Ττ Ττ/Δδ Υυ Φφ Χχ Ψψ Ψψ>

An alphabet based primarily on the Roman script (with the addition of letters from the Greek script) was proposed by linguist Alexander Borg and was officially adopted by the community in a ceremony in 2007 (Armostis, 2020; Karyolemou and Armostis, 2020). This alphabet was subsequently used in community contexts, including the 'Maronite Press' newspaper. Borg, being a speaker of Maltese, a peripheral variety whose alphabet uses the Roman script, modelled his Cypriot Arabic alphabet on the Maltese one, hence its primarily Roman character. Borg's alphabet, as presented in Borg (2008), is replicated below:

- (2) <Aa, Bb, Cc, Ċċ, Dd, Δδ, Ee, Ff, Gġ [or Gg], Xx, Ii, Jj, Kk, Ll, Mm, Nn, Oo, Pp, Rr, Ss, Šš, Tt, Θθ, Uu, Vv, Ww, Yy, Zz>¹

1. Arguably, 'Gġ [or Gg]' found in Borg (2008) was a typo; one would have expected 'Ġġ [or Gg]'.

However, the practical application of this alphabet displayed considerable graphematic variation (Armstrong, 2020), a pattern also observed in the orthographic transcription of the oral corpus compiled within the revitalisation programme (Armstrong, 2025). The main reason for this inconsistency was that only the alphabetisation component of the graphisation process was provided, i.e., the creation and adoption of the alphabet, whereas the second essential component, namely a coherent set of orthographic rules governing its use, was largely absent, beyond basic graphophonemic correspondences. For instance, in the leaflet (Borg, 2008) in circulation within the community about his proposed alphabet, it was specified that the letters <č> and <ġ> represent the affricate consonants [tʃ] and [dʒ] respectively, but only in loanwords. However, it was not explicitly stated how the same affricates would be spelled in inherited words. In the sample text provided in the leaflet for how the alphabet could be applied, it could be found that, in inherited words, these affricates were spelled with the digraphs <tʃ> and <tj> respectively. Additional limiting factors in the application of Borg's proposal included the scarcity and fragmentation of publicly available documentation about the alphabet, the absence of established letter names, and the fact that speakers had not been systematically consulted regarding their preferences.

More generally, the creation of a signary and orthography for previously unwritten or underwritten varieties is often pursued as a largely top-down process with limited speaker participation. However, as speakers' preferences often take precedence over a linguistically ideal orthography (Cahill, 2014; Cahill and Karan, 2008), the research programme opted for a full graphisation process that would involve community members in reaching a signary and orthography proposal, which would be put to test with native speakers. Hence, one of the primary objectives of the Cypriot Arabic documentation and revitalisation programme was the development not only of a revised and fully specified signary, building on and reassessing Borg's earlier alphabetic proposal, but also of a comprehensive orthographic framework (Armstrong, 2025). With respect to the signary, this involved evaluating whether the existing alphabet adequately represented all phonemic contrasts of the language and making decisions concerning alphabetical order, letter names, and the inventory of punctuation and auxiliary symbols. With respect to orthography, the aim was to formulate explicit rules covering graphophonemic correspondences and their balance with morphographic considerations, orthographic depth, the marking of suprasegmental features such as stress and consonant gemination, capitalisation, orthographic word division and hyphenation, contraction, loanword spelling, and punctuation practices (*ibid.*; Coulmas, 1996, p. 379–380). Establishing such conventions was necessary both for ensuring consistency in corpus transcription and for supporting the production of

teaching materials for younger and adult learners within the broader framework of acquisition planning.

1.3. Research Motivation

In designing a writing system from a linguistically informed perspective, a central question concerns the appropriate level of representation: whether priority should be given to surface phonetic forms or to underlying phonemic and morphophonological structure, and the extent to which morphological or etymological information should be encoded. Although linguists typically aim at maximising structural adequacy and internal consistency, linguistic soundness alone does not guarantee the success of a newly proposed writing system. Even a theoretically optimal system cannot be considered ‘ideal’ if it is not acceptable to its intended users, usable in everyday practice, and teachable within realistic educational settings. Research on orthography development has repeatedly shown that acceptability, usability, and learnability, alongside educational, sociopolitical, and practical production factors, are decisive for successful implementation (cf. Cahill, 2014; Cahill and Karan, 2008). These factors include, among others, the transfer of graphetic and graphematic conventions from familiar writing systems (Malone, 2004; Smalley, 1964), as well as technological support such as Unicode encoding, keyboard and font availability, and spellchecking tools. For this reason, proposals developed exclusively on theoretical linguistic grounds must be empirically tested with native speakers, and user preferences and attitudes must be taken into account before a system is finalised (Armostis, 2025; Karan, 2013).

1.3.1. *Morphophonological Processes Under Investigation*

A writing system is analysed as an ordered language-script pair (Justeson, 1976; Weingarten, 2011) consisting of a signary and associated graphematic correspondences between the language and its written form (Daniels and Bright, 1996, p. xlv), part of which may be codified in an orthographic module. A central design issue is representational depth, namely whether spellings should be phonetic or morphophonological.

As a Semitic language, Cypriot Arabic exhibits templatic morphology characterised by the use of consonantal roots, i.e., abstract discontinuous consonant sequences which, through the insertion of transfixes (vowels and additional consonants within the root), yield inflected and derived word forms. Depending on the morphophonological environment, the consonants of a root may undergo systematic (morpho)phonological processes. For example, the root |ktp| ‘write’ appears unchanged in (3a),

partially modified in (3b), and doubly modified in (3c), due to a Cypriot Arabic rule whereby plosive clusters undergo dissimilation with respect to continuancy (Armostis, 2020):

- (3) The root |ktp|
- a. /'kɪtɐp/ → ['cɪtɐp] 'he wrote'
 - b. /'pɛktɔp/ → ['pɛxtɔp] 'I write'
 - c. /ktuɪp/ → [xtuɪt] 'I wrote'

The question that arises is whether the forms in (3b) and (3c) should be written morphophonologically as <paktɔp> and <ktuɪp> or phonetically as <paɪtɔp> and <ɪtuɪt>. The first option gives the writing system greater *orthographic depth/opacity* (Coulmas, 2003, p. 213–215; Rogers, 2005, p. 275), in that different phonetic allomorphs of the same morpheme are written in the same way, something that makes their connection apparent in writing. The second option increases *orthographic transparency*, in that it more faithfully represents the surface pronunciation.

A second phenomenon of interest concerns the spelling of the definite article. The underlying form of the definite article in the nominative, accusative, and dative is the consonant /l/, which phonetically assimilates fully to the consonant of the following word when that word begins with a singleton consonant (rather than a geminate or a consonant cluster), with the exception of pharyngeal [ʕ], which never occurs as a geminate in Cypriot Arabic (Christodoulou, Armostis, and Karyolemou, 2021):

- (4) The definite article /l/
- a. /l#'ɪfɛʔ/ → ['ɪfɛʔ] 'the snake'
 - b. /l#xɛrɪ/ → [xɛrɪ] 'the heat'
 - c. /l#kɛ'fɛ/ → [k^h:ɛ'fɛ] 'the coffee'²
 - d. /l#'pɛpɛ/ → ['b:ɛpɛ] 'the door'

In examples (4b)–(4d), the phonetic realisation differs from the underlying morphophonological form, and the question arises whether the

2. In Cypriot Arabic, the geminate counterparts of the voiceless occlusives [p], [t], [k], [c], [tʃ] are enhanced by aspiration, i.e., [p^hː], [t^hː], [k^hː], [c^hː], [tʃ^hː]. This phenomenon is an influence from Cypriot Greek, in which these occlusives are also enhanced by aspiration when geminate (Armostis and Karyolemou, 2023). It should be noted that Cypriot Arabic additionally exhibits two geminate voiced plosives, namely [bː] and [dː] (which do not exist in Cypriot Greek), with no singleton counterparts (singleton [b], [d], and [g] do appear, but only as allophones of /p/, /t/, and /k/ respectively when immediately preceded by a voiced consonant: e.g., /pi'ntuɛ/ → [pi'nduɛ] 'little girl'; cf. [pnɛt] 'daughters').

article should be written as pronounced in each case or whether it should always be written unchanged in its underlying form. For instance, (4b) may be represented morphophonologically as <l-χarr> or phonetically as <χ-χarr> (or <χχarr>, if the hyphen is not used). If (4c) is written more phonetically as <kkafé> (without the hyphen), the phrase ‘the coffee’ becomes homographic with <kkafé> ‘cafeteria’, whereas a morphophonological spelling <lkafé> (and even better: <l-kafé>) keeps the two forms distinct and encodes additional morphographic information (cf. Rogers, 2005, p. 275). The situation is even more complex in (4d), where assimilation of /l/ to /p/ obligatorily yields [b:] rather than [p^h]; thus the phonetic spelling <bbape> / <b-bape> departs even further from the underlying form (and from the form without the article, i.e. <pape>) than in the previous examples, whereas the morphophonological spelling <lpape> / <l-pape> may facilitate visual word recognition (Armostis, 2020). Also, even though the assimilation of /l/ to the coronal /t/ typically yields a voiceless geminate [t^h] (e.g., /l#tik/ → [t^hik] ‘the rooster’), in a couple of lexical items the result of this assimilation is rather a voiced [d:]: for instance, /l#tist/ → [d:ist] ‘the cauldron’. The reasons for both the regular /l#p/ → [b:] assimilation and the exceptional cases, such as /l#tist/ → [d:ist], can be found in the diachrony of the language (Borg, 1985). These irregular cases also pose the same problems as the /l#p/ → [b:] assimilation for the written representation of such phrases: should it be written phonologically as <l-tist> (cf. the form without the article: <tist>) or phonetically as <d-dist>?

Even in case (4a), where phonological and phonetic form do not differ, the issue still arises of how the article should be written: (a) attached to the following word, i.e., <lcafa>; (b) as a separate word, i.e., <l cafa>; (c) with an apostrophe, i.e., <l'cafa> (a practice attested in usage); or (d) joined with a hyphen, i.e., <l-cafa>. Writing the definite article /l/ as a separate word results in a word spelled without a vowel letter (i.e., <l cafa>), which may appear unusual, since phonological words are expected to contain a syllabic nucleus. Writing it attached (i.e., <lcafa>) conforms to phonotactic structure (the cluster /lʃ/ is allowed word-initially: e.g., [lʃɛft] ‘I played’), but is not morphologically transparent, as it does not indicate that the article is a separate morpheme. The apostrophe is not optimal (i.e., <l'cafa> or <l' cafa>), since it conventionally signals vowel elision, which is not the case here. Hyphenation indicates that the sequence forms one phonological word but two morphological units. In our proposal, hyphenation was adopted (as also done by Borg), with /l/ written invariantly as <l-> in all contexts: e.g., <l-cafa>, <l-χar>, <l-kafé>, <l-pape>, <l-tist>. These various ways of spelling the definite article were included in the test design of the present study.

A third phenomenon investigated in the present study concerns a small set of voiced ‘fricative + plosive’ clusters, namely [ðb], [ɣb], and [ɣd]. These clusters present a particularly interesting challenge for or-

TABLE 1. Examples of Cypriot Arabic words containing voiced 'fricative + plosive' clusters (adjusted from Borg, 1985, p. 15)

surface form	Borg's underlying representa- tion	reconstructed historical form	etymology
[ɣbir] 'big (m.)'	← /ɣpir/	< *gbi:r	< (Old Arabic) kabi:r
[rɐɣbɐ] 'neck'	← /'rɐɣpɐ/	< *rɐgbɐ	< (Old Arabic) raqaba
[rɐɣbɐ] 'riding (f.)'	← /'rɐɣpɐ/	< *rɐgbɐ	< (Old Arabic) ra:kiba
[ðbɐxt] 'I slaughtered'	← /tɐxt/	< *dbɐxt	< (Old Arabic) ðabaht
[pɐ'ðbun] 'river'	← /pɐ'θɐpun/	< *pɐ'tmun	< (Cypriot Greek) pɔtɐ'mɔn
[ɣdiɟ] 'horse'	← /ɣtiɟ/	< *gdi:ɟ	< (Old Arabic) kadi:ɟ
[mɔɣdɐ] 'stove'	← /'mɔɣtɐ/	< *mogde	< (Old Arabic) mawqida

thographic representation, because their synchronic phonological status is far less straightforward than that of the processes discussed above: unlike the alternations involving continuancy dissimilation or full assimilation of the definite article /l/, which can be analysed in terms of productive morphophonological processes, the voiced clusters appear to be the result of historical sound change and have a highly restricted lexical distribution. Such items are presented in Table 1, adjusted from Borg (ibid., p. 15).

Borg (ibid., pp. 15–16) postulated that, synchronically, Cypriot Arabic exhibits a voiced plosive in the underlying representation of such clusters, as, historically, it was the second member of the cluster that was phonemically voiced: e.g., [ɣbir] 'big (m.)' ultimately goes back to Old Arabic /kabi:r/. However, postulating voiced plosives in the phonemic form is problematic, as Cypriot Arabic does not have an underlying voicing contrast for singleton plosives (see footnote 2): it exhibits only a series of voiceless plosive phonemes, namely /p/, /t/, /k/, which surface as voiced when immediately following another voiced consonant (ibid., p. 12). The alternative, albeit not historically motivated, would be to assume that, synchronically, the first member of the cluster is underlyingly voiced, to which the second member assimilates. However, this solution is also problematic, as Borg (ibid., p. 36–38) argues that Cypriot Arabic does not synchronically exhibit the Old Arabic /ɣ/ phoneme due to it having merged with /ʕ/ in Cypriot Arabic (e.g., Old Arabic /ɣalaq/ is the origin of Cypriot Arabic [ʕilek] 'he shut'). Consequently, it is not al-

ways evident which segment should be regarded as underlyingly voiced from a synchronic perspective, nor is it clear whether the surface clusters should be analysed as the output of an active phonological process converting the voiceless cluster into a voiced one, as such a rule would be active only if historically motivated: e.g. it would apply for /xtif/ → [ɣdif] ‘horse’ (cf. Old Arabic /kadi:f/), but not for /ktupt/ → [xtuft] (cf. Old Arabic root |tkb|). This conundrum led Borg (1985) to represent the voiced ‘fricative + plosive’ clusters underlyingly with voiceless phonemes, which however have exceptionally received voicing in certain lexical items due to historical reasons, e.g., /x̥pir/.

This theoretical opacity of the synchronic underlying form of such clusters has consequences on their written rendition. If a more surface representation is adopted, then spellings such as <ɣbir>, <paðbún>, <ɣdiş> etc., are straightforward, as their phonetic form unambiguously contains a voiced fricative followed by a voiced plosive; but, if a deeper representation is adopted, then, apart from the fact that from a theoretical perspective the phonological form is unclear, an additional issue would be how deep the written representation should be with regards to the morphophonology or even etymology of a given item: e.g., would a deeper spelling for [ɣdif] be <χtiş>, <χdiş>, <gdiş>, or <kdiş>? The phenomenon therefore provides an additional test case for examining speakers’ preferences regarding orthographic depth and the extent to which historical, morphophonological, or phonetic information should be represented in the writing system.

1.3.2. *The Proposal*

The proposal of the research programme that was to be tested was the following:

Regarding the signary, it was developed by making decisions about the alphabet, punctuation (e.g., using the Roman question mark <?> rather than the Greek one <;>), numerals (Western Arabic numerals were selected), decimal separator etc. For the purposes of the test of the current study, only the alphabet part of the signary was presented to the participants. The test alphabet was a modification of the one created by Borg:

- (5) <Aa Bb Cc Dd Δδ Ee Ff Gg Γγ Ii Jj Kk Ll Mm Nn Oo Pp Θθ Rr Ss Şş Tt Uu Vv Ww Xx Yy Zz>

This version of the alphabet was not meant as a finalised proposal, but was rather put together to be tested in order to be modified should this investigation yield serious difficulties with certain aspects thereof.

Regarding the orthographic component, certain spelling conventions were included in the materials to be presented to the participants of this

study. Apart from basic graphophonemic correspondences, the proposal contained rules for marking stress, writing diphthongs, affricates, and voiced plosives. Regarding voiced plosives, the proposed rule presented was that they are written with and <d> only when geminate: e.g., <Rabbi> for [ˈræbːi] ‘my God’; <ṣadda> for [ˈʃædːæ] ‘he tightened it (f. object)’. The rule for phonetically singleton voiced plosives [b], [d], and [g] was that they always occur after other voiced consonants and are represented with the letters for voiceless consonants, i.e., with <p>, <t>, <k> respectively: e.g., <mkass> for [mɡæsː] ‘scissors’. The only exception to this rule was the representation of the voiced ‘fricative + plosive’ clusters: in our proposal, we opted for the unambiguous surface representation, i.e., <ḍb>, <ḡb>, <ḡd>, for example, <ḍbaḡt>, <ḡbir>, <ḡdiṣ>. Regarding the definite article and continuancy assimilation, no explicit proposal was made; the reason for this was for us to be able to observe how native speakers would represent the two phenomena without being influenced by our preferences.

1.3.3. *Research Questions*

Based on the theoretical and practical considerations outlined above, the present study formulated the following research questions concerning the graphisation of Cypriot Arabic:

1. Which script (or scripts) should form the basis of the Cypriot Arabic signary? More specifically, given the literacy backgrounds of the community and their familiarity with particular writing traditions, which graphematic resources are most appropriate for the development of a usable and sustainable writing system?
2. Which level of linguistic representation should a linguistically well-motivated orthography primarily respect? Should it represent:
 - (a) surface phonetic forms, resulting in a more transparent orthography; or
 - (b) underlying morphophonological representations, resulting in greater orthographic depth (and increased morphography)?
 - (c) And should the selected level of representation (surface vs underlying) be uniformly applied to all phenomena of the language?
3. How acceptable, usable, and teachable is the proposed writing system as evaluated by members of the speech community themselves?

Together, these questions link structural linguistic considerations with user-centred criteria and treat graphisation both as a technical design task and as a socially embedded process.

2. Methodology

To answer the above research questions, an empirical study was designed, as presented below.

2.1. Materials

Data were collected with the use of a questionnaire. The questionnaire was administered in printed form, while the instructions were presented on a laptop screen using MS PowerPoint. Each test word or phrase was presented aurally to the participants through headphones.

2.1.1. *The Questionnaire*

The questionnaire consisted of three main parts: (1) a pre-test; (2) a brief teaching session presenting the basic rules of the proposed writing system along with a questionnaire for evaluating the proposal; and (3) a post-test. Its overall design aimed to elicit participants' spontaneous spelling practices, measure their preferences when confronted with alternative spellings, and gauge their evaluation of the proposed alphabet and orthography after minimal instruction.

The pre-test comprised two tasks. In the first task, participants listened to recorded Cypriot Arabic words and short phrases and were asked to write each item down by hand in whatever way they considered appropriate. This task was intended to capture uninstructed spelling strategies and script choices. In the second task, participants again listened to recorded items, but this time a forced-choice format was followed, whereby they were asked to select their preferred spelling among a set of proposed alternatives. These alternatives reflected different graphematic choices (e.g., more phonetic vs more morphophonological spellings, different boundary-marking options) etc.

The second part of the questionnaire consisted of a very brief instructional presentation of the proposed writing system, including the alphabet (signary) and its principal orthographic conventions. Immediately following this introduction, participants completed the second part of the questionnaire evaluating the proposed system. Using five-point Lickert-scaled responses, they were asked to indicate: how much they liked the orthographic rules presented to them; the degree to which they agreed with these rules; to what extent they would personally follow them when writing in Cypriot Arabic; to what extent they believed that all users should follow these rules; how logical they found the rules; and how easy they considered them to be learnt.

TABLE 2. The target determiner phrases used in each part of the test for the investigation of the written rendition of full assimilation of the definite article

Part	target phrase	Gloss
pre-test (a)	/l#'vɛrɛk/ → ['v:ɛrɛk]	the books
	/l#tik/ → [t ^h :ik]	the rooster
pre-test (b)	/l#'ʒɛʒʒɛ/ → ['ʒ:ɛʒ:ɛ]	the hen
	/l#kɛ'ssis/ → [k ^h :ɛ's:is]	the priest
post-test (a)	/l#xɛpl/ → [x:ɛpl]	the rope
	/l#tist/ → [dɪst]	the cauldron
post-test (b)	/l#'pɛpɛ/ → ['b:ɛpɛ]	the door
	/l#'θɛɥp/ → ['θ:ɛɥp]	the shirt
	/l#pint/ → [b:ɪnt]	the daughter

The post-test also consisted of two tasks, presented in the reverse order compared to the pre-test: participants first completed a forced-choice spelling task based on recorded words, again selecting among proposed spellings; they then performed a final handwriting task, writing down the recorded items. This reversed ordering allowed for comparison of preferences and production before and after exposure to the proposed system and its rules.

2.1.2. *Target Words*

The words and phrases used in the study were embedded as audio files in the PowerPoint file used for the test. To create this material, EZ, a 45-year-old male native speaker of Cypriot Arabic, was recorded in a quiet room. EZ was familiar with the proposed writing system, as he was a member of the research project for the documentation and preservation of Cypriot Arabic. The words and phrases under investigation were presented to the speaker in writing. EZ was instructed to pronounce the words and phrases clearly and to avoid 'list intonation' by pronouncing each one separately.

The present study formed part of a wider investigation, the design of which included a wide range of phenomena to be examined through a total of 129 words and phrases serving as the test material. The current study though was confined to the investigation of three specific phenomena: (1) the complete post-lexical assimilation of the definite article to the following word-initial consonant within a determiner phrase (Table 2); (2) continuancy dissimilation of two-plosive clusters (Table 3); and (3) the voicing of 'fricative + plosive' clusters (Table 4).

TABLE 3. The target words used in each part of the test for the investigation of the written rendition of continuancy dissimilation of two-plosive clusters

Part	target word	Gloss
pre-test (a)	/pkum/ → [fkum]	I get up
	/ptə'min/ → [ftə'min]	rivers
pre-test (b)	/sruxt/ → [sruxt]	I stole
	/tlupt/ → [tluft]	I asked
post-test (a)	/sukt/ → [suxt]	I ploughed
	/'tkurə/ → ['θkurə]	males
post-test (b)	/ktilt/ → [xtilt]	I killed
	/ktupt/ → [xtuft]	I wrote

TABLE 4. The target words used in each part of the test for the investigation of the written rendition of the voiced clusters [ðb], [ɣb], and [ɣd]

Part	target word	Gloss
pre-test (a)	[ðbæxt]	I slaughtered
	[ɣbir]	big (male-singular)
pre-test (b)	[ðbu'næt]	straw (plural)
	['reybe]	on horseback (female adjective)
post-test (a)	[ɣdif]	horse
	[pə'ðbun]	river
post-test (b)	['mɔyde]	oven
	['reybe]	neck

2.2. Participants

Twenty-two participants (10 men and 12 women) took part in the study, ranging in age from 39 to 72 (mean: 52.3; SD: 8.19). Two participants had received only primary education, 12 secondary education, and 8 tertiary education. All participants had Cypriot Arabic as their first language and were bilingual with Cypriot Greek.

2.3. Procedure

Before the main part of the study began, participants filled out a questionnaire about their demographics, as well as about whether and to what extent they use written Cypriot Arabic, and also about whether they know other languages and to what degree.

They were then given the first printed section of the questionnaire with the instruction to freely write down with a pen the word/phrase recordings they would hear through headphones connected to the laptop displaying the PowerPoint show. It was emphasised to them that this was not a test, as there is no right or wrong answer; on the contrary, it

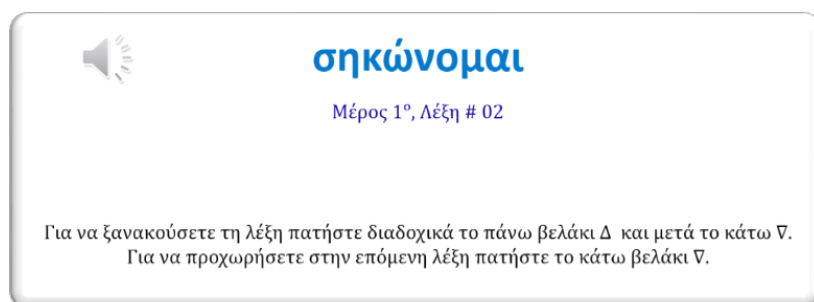


FIG. 1. A slide from the PowerPoint presentation used during the test

was stated that the researchers wanted to learn from the participants regarding their preferences in rendering Cypriot Arabic in writing. Each PowerPoint slide displayed the serial number of the word being played as well as its translation into Standard Modern Greek. In addition, instructions were provided on how to replay a word if they wished (see Figure 1). As each slide appeared, the corresponding audio clip began playing automatically.

2.4. Analysis

In each part of the test, the rates of written representation of the underlying morphophonological level vs the surface phonetic level of the three phenomena under examination (i.e., plosive cluster continuancy dissimilation, full assimilation of the definite article to the following consonant, and the voicing of ‘fricative + plosive’ clusters) were measured. Regarding the evaluation of the proposed writing system, as responses were given on a five-point Likert scale (1–5), the appropriate measures of central tendency and dispersion for such ordinal scale were calculated, namely, the median (*Mdn*), mode (*Mo*), and range.

3. Results

3.1. Evaluation of the Proposed System

As shown in Figure 2, participants rated the proposed writing system as fairly logical (*Mdn* = 4; *Mo* = 4) and stated that they somewhat liked it and would probably use it (*Mdn* = 3.5; *Mo* = 4 for both). They did not seem to have a strong opinion on whether they agreed with this proposed system (*Mdn* = 3; *Mo* = 3), and there was not much variation in their responses (range = 2–4).

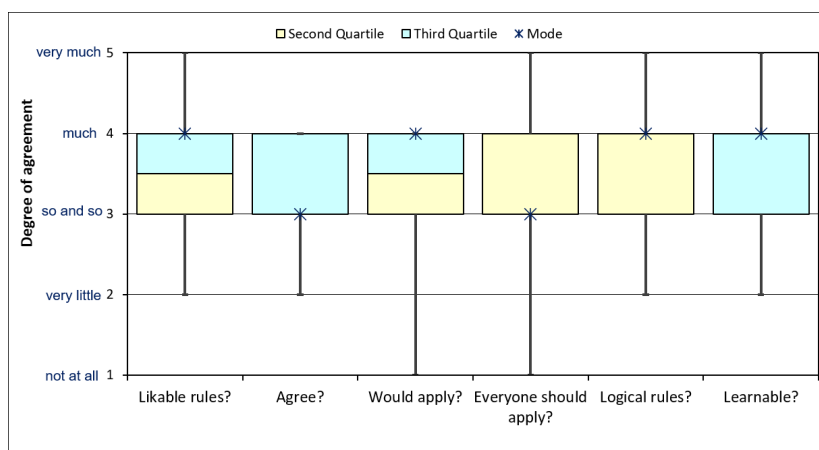


FIG. 2. Boxplots of evaluation of the proposed system

Furthermore, they did not have a strong opinion on how easy it is to learn this system, although the most frequent answer was that it is easy to learn ($Mdn = 3$; $Mo = 4$). Finally, they seem to agree that the proposed system should be implemented by everyone, although the most frequent answer was that they did not have a strong opinion on the matter ($Mdn = 4$; $Mo = 3$), while there was a wide range of responses (range = 1–5).

3.2. Graphematic Choices

Figure 3 shows the graphematic choices of the participants in the parts in which they could write down the target words and phrases by hand, i.e. in the 'pre-test (a)' and the 'post-test (b)'. Regarding script choice, they would initially write in entirely the Greek script at a rate of 34.3% at the very start of the test; this percentage dropped to 14.9% at the very last part of the test. Regarding the use of the proposed alphabet and spelling conventions, their choices followed the proposed system (even before it was introduced to them) at a rate of 10.3% at the beginning of the test (this number counted only segmental graphemes ignoring the use of stress marks); at the end of the test, i.e. after they had been briefly taught the proposed conventions, this use of the proposed system increased to 18.2%. In the vast majority of the cases, they would use mixed conventions, i.e. they would oscillate between the proposed and ad hoc conventions, or they would use completely ad hoc conventions sometimes influenced by the writing systems of foreign languages they had learnt through education, such as English, French or Italian; as expected, no use of the Arabic script emerged. The percentage of such mixed use

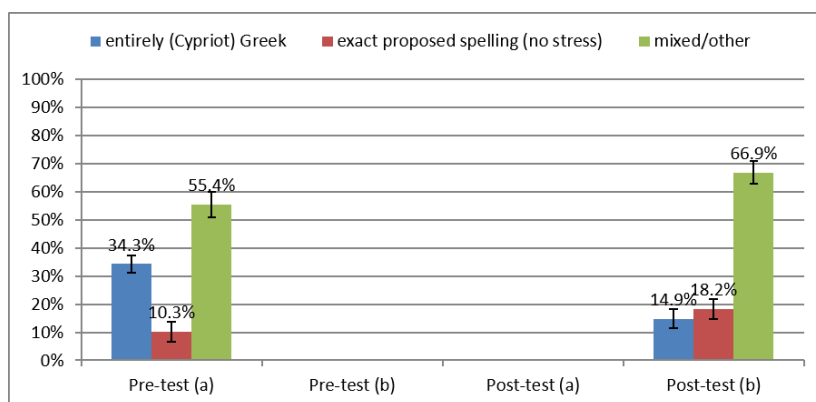


FIG. 3. Percentages of use of the proposed orthography vs. alternatives in the non-forced-choice parts of the test. [Note: In this and the following graphs, error bars represent standard error for the mean.]

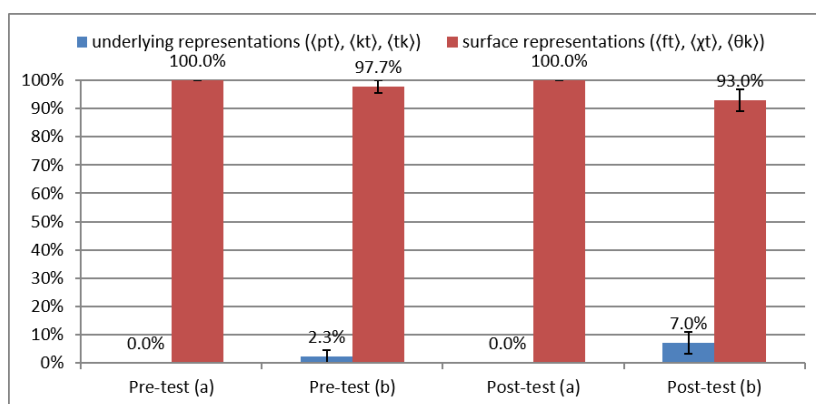


FIG. 4. Percentages of underlying vs surface spelling of two-plosive clusters

was higher at the end of the test compared to what the situation was at the beginning.

With regards to the written representation of two-consonant clusters consisting phonologically of two plosives, a clear preference was shown ($M = 98.1\%$; $SD = 13.7\%$) for representing the surface phonetic result of the continuancy dissimilation process (e.g., [xt] rendered as <χt>) rather than the underlying form (e.g., /kt/ as <kt>). As shown in Figure 4, phonetic spelling was the most prevalent in all parts of the test.

In the case of the definite article, phonetic spelling was also generally preferred (e.g. [t^h:ik] rendered as <t-tik>) over the morphophonological one (e.g., /l#tik/ as <l-tik>), but not quite as clearly ($M = 69.8\%$; $SD =$

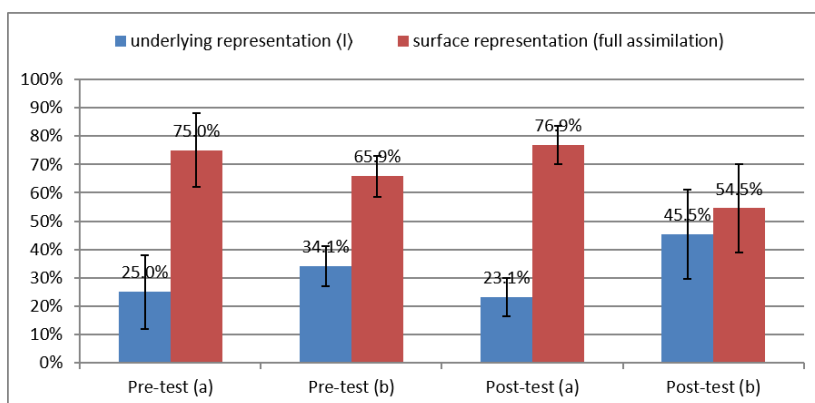


FIG. 5. Percentages of underlying vs surface spelling of the definite article /l/

46.1%). As shown in Figure 5, the morphophonological spelling <l> was used between 23.1% and 45.5%, with the highest percentages recorded in the final part of the test.

As for the rendering of the word boundary between the definite article and the noun it modifies, in the pre-test (a), where participants wrote freely, the main tendency was to write them together as a single word (e.g., <ltik> or <ttik>). In post-test (b), writing with the hyphen (e.g., <l-tik>) became the most prevalent representation. As shown in Figure 6, the other spellings declined in favour of the hyphenated spelling; the only exception was the spelling without the article, which consistently had high percentages in both parts. The use of a simple typographical space between the two words (e.g., <l tik>) was not used in either part. It should be noted that Figure 6 does not include the forced-choice tasks, as in those ones the options given always used the hyphen (e.g., <l-kassís> vs <k-kassís>).

The results regarding the representation of the voiced 'fricative + plosive' clusters [ðb], [ɣb], and [ɣd], are presented in Figure 7. In both the initial handwriting task and the following forced-choice task of the pre-test, the representation of only the fricative as voiced and the plosive as voiceless (e.g., <ðpayt>, <ɣpir>, and <ɣtiş>) was clearly preferred. In the forced-choice task of the post-test immediately after the presentation of the proposed orthographic rules, there was a reversal of preferences, with the proposed surface representations (e.g., <ðbayt>, <ɣbir>, and <ɣdiş>) being favoured. In the final handwriting task though, they reverted to the preference of the 'voiced fricative + voiceless plosive' representation. It should be noted that spellings representing both members of the cluster as voiceless were minimally selected only in the forced-choice tasks.

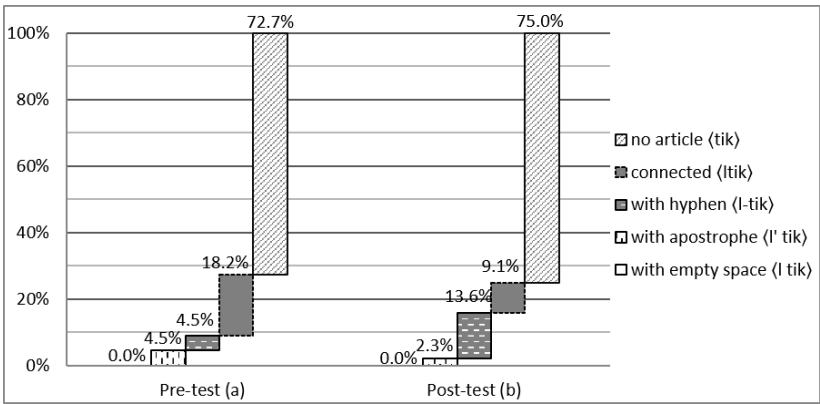


FIG. 6. Percentages of the different ways in which the definite article /l/ was linked in writing to the following noun per the two non-forced-choice parts of the test

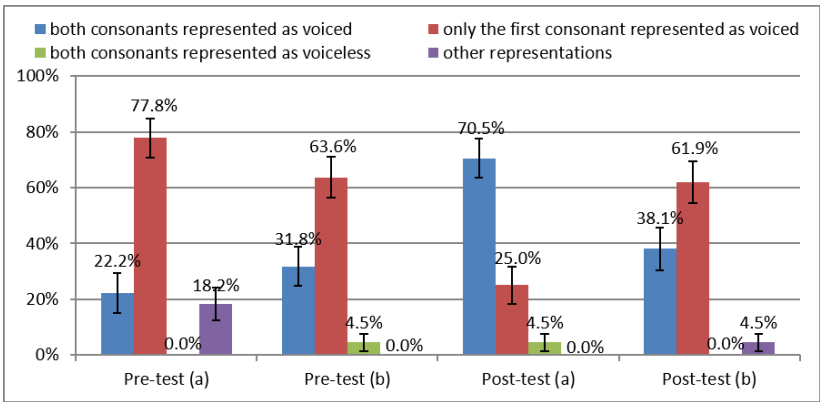


FIG. 7. Percentages of various ways in which the [ðb], [ɣd], [ɣb] clusters were represented in writing

4. Discussion

The present study set out to investigate three interconnected questions regarding the graphisation of Cypriot Arabic: which script should form the basis of the proposed signary and associated orthography, which linguistic principles should govern the orthography, and to what extent native speakers find the proposal logical, learnable, and likeable. The results provide insights into all three issues and demonstrate that graphisation cannot be reduced to solely the pursuit of a linguistically ideal orthography; rather, as argued in the language-planning and

grapholinguistic literature (Cahill, 2014; Cahill and Karan, 2008; Haarmann, 2006), the successful implementation of a proposed orthography depends on balancing structural adequacy with acceptability, usability, and learnability.

With regard to the choice of script, the results confirm the importance of speakers' literacy backgrounds. No participant made use of the Arabic script, despite the fact that Cypriot Arabic is an Arabic variety. This finding is unsurprising given that community members are literate in Standard Modern Greek and familiar with the Roman script through foreign-language education, but generally have no literacy in Modern Standard Arabic (with the exception of a few who learnt it as a foreign language). The spontaneous use of Greek-script spellings at the beginning of the test and the frequent adoption of mixed conventions drawing on both Greek- and Roman-script practices further illustrate the tendency of writers to transfer graphematic conventions from familiar writing systems. At the same time, the increased use of the proposed alphabet following even a brief instructional intervention suggests that script choice is not determined solely by prior literacy experience, but can be influenced through teaching and standardisation efforts.

The most theoretically interesting findings concern the question of orthographic depth and the level of linguistic representation that a writing system should respect. As discussed earlier, a central issue in orthography design is whether spellings should (consistently) privilege phonetic transparency vs. morphophonological deeper representations. The results demonstrate that the answer cannot be reduced to a simple choice between a shallow and a deep orthography. Rather, speakers showed different preferences for different types of morphophonological phenomena.

The strongest preference observed in the study concerned the representation of continuancy dissimilation in dyadic plosive clusters. Participants overwhelmingly preferred spellings reflecting the surface phonetic outcome of the process rather than the underlying morphophonological representation. Thus, forms such as /ktuɸt/ → [xtuɸt] were overwhelmingly written as <xtuɸt> rather than <ktuɸt>. This finding can be interpreted within the framework of Lexical Phonology (Kiparsky, 1982). Lexical phonological rules are integrated into speakers' lexical representations and are often sensitive to morphological structure. Consequently, the outputs of such rules may acquire psychological reality for speakers and be perceived as the 'actual' form of a word. In Cypriot Arabic, continuancy dissimilation in plosive clusters exhibits precisely such characteristics and is not merely an automatic (i.e., post-lexical) phonological process: more precisely, the dissimilation rule is suppressed in lexical items such as ['ʔeptə] '(Santa) Maura', ['piptəl] 'I change', [ptar] 'seed'; as Borg (1985, p. 18) observed, there is even a minimal pair, namely [ptilt] 'I changed' (root: |ptl|) vs [ftilt] 'I twisted' (root: |ftl|), with the for-

mer suspending the dissimilation rule arguably for remaining distinct in form from the latter. The existence of lexical exceptions, the morphological conditioning of the process, and also the functional load of the process due to the existence of the above-mentioned minimal pair, suggest that speakers store and recognise the resulting forms as part of their lexical knowledge. It is therefore unsurprising that they strongly preferred surface representations such as <ɣtuft> and <paytop>.

A different pattern emerged in the case of the definite article. Although phonetic spellings were generally preferred throughout, the preference was substantially weaker, and morphophonological spellings increased once this option was presented to the participants in pre-test (a) and post-test (b). Importantly, as mentioned above, in the presentation of the proposed rules, no explicit preference for the writing of the definite article was introduced to the participants; this means that the substantially increased percentage of selecting the orthographically deep option in post-test (b) may be attributed to it being a rather intuitive option. This interpretation may be supported by the fact that in Borg's (2008) leaflet for his proposed alphabet, the phonetic level of representation was used in writing the definite article (e.g., <t-tarp> 'the road'), as is the case in Maltese (e.g., <it-tifel> 'the child'); if Cypriot Arabic speakers had already been familiar with Borg's (2008) conventions, then their increased percentages of using a deep representation by the end of the test is quite revealing. This result is likewise compatible with a Lexical Phonology interpretation: assimilation of the definite article /l/ occurs automatically across word boundaries and therefore constitutes a post-lexical process. Unlike lexical rules, post-lexical rules are typically exceptionless,³ automatic, and less accessible to conscious awareness, because they apply after lexical representations have been assembled. Consequently, speakers may be more willing to preserve the underlying representation of the article in writing, especially when doing so enhances morphological transparency. The increased acceptance of forms such as <l-tik> by the end of the test therefore suggests that speakers recognise the value of maintaining a fixed image for the writing of this grammatical morpheme, despite the fact that its phonetic realisation varies, exactly because the phonetic outcome is completely predictable by the phonological rules of the language.

A somewhat different pattern emerged in the case of the voiced 'fricative + plosive' clusters [ðb], [yb], and [yd]. Unlike the phenomena discussed above, these clusters do not correspond straightforwardly to an active morphophonological process. As outlined in section 1.3.1, their

3. Even the /l#tist/ → [dɪst] case, albeit an exception to the /l#t/ → [tʰ:] rule, it is not really a suspension of the full assimilation rule, but rather an application thereof with an unusual assimilation outcome; as such, it remains an automatic process and arguably below the level of conscious awareness.

synchronic status is theoretically opaque under Borg's (1985) analysis: although they are phonetically realised with two voiced obstruents, their historical origins differ from item to item, and it is not clear whether their synchronic written representation should preserve information about etymology, an abstract phonological representation, or the surface phonetic form. Consequently, if we follow Borg's (1985) analysis, the phenomenon occupies an intermediate position between synchronically productive morphophonology and lexically restricted historical residue.

The results indicate that participants were generally reluctant to represent both members of the cluster as voiced. In both tasks of the pre-test and in the final handwriting task, they clearly preferred spellings such as <γpir>, <δpaxt>, and <γtiş> over the proposed surface spellings <γbir>, <δbaxt>, and <γdiş>. Only immediately after presentation of the proposed orthographic rules did participants temporarily reverse this preference and favour the fully voiced representations. This suggests that, while speakers found the proposed convention understandable and were willing to adopt it when explicitly prompted, it was not their most intuitive solution when writing spontaneously. An explanation for this preference may lie in the phonological structure of Cypriot Arabic itself. As discussed earlier, the language does not possess an underlying voicing contrast among singleton plosives; instead, voiced plosives are normally predictable allophones of /p/, /t/, and /k/ occurring after voiced consonants (Borg, 1985). Participants therefore appear to have extended this general graphematic principle to the exceptional voiced clusters as well. By writing forms such as <γpir> rather than <γbir>, they preserved the salient surface voicing of the fricative while simultaneously maintaining the conventional representation of the plosive with a grapheme corresponding to the underlying voiceless phoneme. At the same time, the near absence of spellings such as <χpir> demonstrates that speakers did not ignore the phonetic voicing of these clusters altogether. Rather than favouring either a fully phonetic representation or a deeper morphophonological-etymological one, they tended to adopt what appeared to be an intermediate solution that captured part of the surface realisation while remaining consistent with broader orthographic principles of the writing system.

This outcome led us to reconsider our proposed <γb>, <δb>, and <γd> spellings and adopt the apparently more intuitive <γp>, <δp>, and <γt> spellings in our final version of the proposed orthography. Part of the reason for this was that we revisited the phonological status of /γ/: Borg (ibid.) did not recognise phonemic status for this segment, as it was thought to be non-contrastive and also historically deriving from a voiceless consonant, such as Old Arabic /k/ or /q/ in clusters with a voiced consonant; however, in preparing our materials for teaching the language, we observed that /γ/ is regularly used in loanwords from Cypriot Greek or Turkish (usually through Cypriot Greek), such as

/ʁe'lliɛ/ 'France', /ɛzʁe'ltin/ 'almost', /ʁur'suz/ 'a jinx male person' etc., something that makes its phonemic status plausible. Moreover, Borg (personal communication), has recently found a minimal pair exhibiting a contrast between /ʁ/ and /x/, namely [ʁdɛʔɛt] 'I touched' vs [xtɛʔɛt] 'I cut (past tense)'; this independent finding provides further support to the idea that /ʁ/ does have functional load and hence should be considered a distinct phoneme in Cypriot Arabic. This justifies our decision to include a letter for it in our proposed alphabet, i.e. <ʁ>, something that had been missing from Borg's alphabet. Accepting phonemic status for the voiced velar fricative may have a further implication: one may argue that the native speakers' preference for the <ʁp>, <ɖp>, and <ʁt> spellings mirrored their intuition for the underlying representation of these clusters, as, synchronically, they may be represented phonemically as straightforward /ʁp/, /ɖp/, and /ʁt/ clusters. This may mean that, even though historically it was the second member of the cluster that was voiced, the clusters may have been reanalysed as consisting of a voiced velar fricative phoneme /ʁ/ or /ɖ/ and a voiceless plosive phoneme /p/ or /t/. If this analysis is accepted, then the synchronic underlying representation of the [ʁb], [ɖb], and [ʁd] clusters becomes considerably more straightforward than under Borg's (1985) account. More importantly, if the fricatives in the cluster are phonemically voiced, then the voicing of the plosives need not be sought in their diachrony, but rather in a phonological rule of regressive voice assimilation of the voiceless plosive to the preceding voiced fricative. Such a process may plausibly be analysed as post-lexical, insofar as the voicing of the plosive follows automatically from the presence of the preceding voiced fricative and does not appear to be lexically conditioned, as, e.g., /ʁt/ → *[ʁt] would be not an acceptable outcome for any lexical item. If this is indeed a post-lexical process, then the interpretation that the <ʁp>, <ɖp>, and <ʁt> spellings clearly represent their phonemic form gains support, as post-lexical processes tend to be represented in their underlying form, as is the case with the definite article. These findings concerning the phonemic status of the voiced velar fricative, together with the results concerning script preferences discussed earlier, as well as other considerations that go beyond the scope of the present study, led to certain modifications of the proposed signary. The final version of the alphabet adopted for the teaching materials is shown below:

- (6) <Aa Bb Cc ʁʁ Dd Δδ Ee Ff Θθ Ii Jj Kk Ll Mm Nn Oo Pp Rr Ss Šš Tt
Uu Vv Ww Xx Yy Zz>

All in all, the results support the view that orthographic design should not be conceived as a binary choice between shallow and deep orthographic representations; instead, different phenomena may warrant different representational strategies. Thus, the emerging Cypriot Arabic

orthography cannot be characterised as either entirely phonetic or entirely morphophonological; rather, it combines shallow and deep orthographic principles depending on the linguistic phenomenon under consideration. The results regarding script preference also confirmed the importance of literacy transfer from familiar writing systems and hence the relevance of sociolinguistic and educational factors in graphisation. The other main outcome of this study pertains to the significance of involving speakers in graphisation processes. Participants generally evaluated the proposed system positively, considering it logical and expressing moderate willingness to use it. More importantly, their responses provided direct evidence regarding which features of the writing system were intuitively acceptable and which required reconsideration. This supports the argument advanced by Karan (2013) and others that writing systems should not be imposed exclusively through top-down expert decisions. Particularly in the case of endangered languages, where community acceptance is crucial for successful implementation, graphisation should be viewed as a collaborative process involving consultation with speakers (Armostis, 2025). Native-speaker preferences offer valuable information not only about attitudes towards particular orthographic options but also about the cognitive salience of linguistic distinctions and the practical usability of competing representations.

The final Cypriot Arabic writing system that emerged from this process therefore reflects both linguistic and social considerations. It is grounded in phonological and morphological analysis, but it is also informed by empirical evidence regarding speaker preferences, literacy practices, and usability. In this sense, the study illustrates how graphisation can function as a form of corpus planning that is simultaneously linguistically informed and community-oriented.

5. Conclusions

The present study examined native speakers' responses to a proposed writing system for Cypriot Arabic and explored how different linguistic phenomena should be represented in the process of graphisation. More specifically, it investigated speakers' preferences regarding script choice, orthographic depth, and the representation of selected morphophonological phenomena, while also assessing the acceptability, usability, and learnability of the proposed system.

The findings demonstrate that the development of a writing system cannot be guided exclusively by theoretical linguistic considerations: although linguistic soundness remains an important criterion, the success of a writing system ultimately depends on whether its intended users find it acceptable, usable, and learnable. The relatively positive attitudes expressed by participants towards the proposed orthography con-

firm the importance of incorporating speaker perspectives into graphisation and orthographic planning, particularly in the context of endangered-language revitalisation. With regard to orthographic depth, the results indicate that a writing system need not be uniformly phonetic or uniformly morphophonological. Rather, different linguistic phenomena may call for different representational solutions: participants tended to favour deeper spellings in cases where the phonological process was post-lexical, while more surface-oriented solutions were preferred in cases whereby the phonological process was lexical. The findings therefore suggest that orthographic depth is not an all-or-nothing property of a writing system but may vary according to the linguistic characteristics of individual phenomena.

More broadly, the study demonstrates the importance of involving speakers in graphisation. Native speakers' preferences provide valuable insight into which orthographic conventions feel intuitive and which create difficulties. Consequently, graphisation should be viewed not as a purely top-down linguistic exercise but as a process in which both linguistic analysis and community perspectives are taken into consideration. The case of Cypriot Arabic graphisation demonstrates the value of empirically evaluating orthographic proposals before their wider implementation. Testing orthographies allows researchers to identify potential difficulties, assess competing design choices, and refine orthographic conventions on the basis of actual speaker behaviour rather than theoretical assumptions alone. Such an approach may be particularly beneficial in the graphisation of endangered languages, where the successful adoption of a writing system is closely linked to broader efforts of language maintenance and revitalisation.

References

- Armotistis, Spyros (2020). "Η γραπτή απόδοση της κυπριακής αραβικής: μια πειραματική προσέγγιση [The written rendition of Cypriot Arabic: an experimental approach]". In: *Έρευνες για την Ελληνική Γλώσσα [Studies on Greek Language]*. Ed. by Elena Ioannidou, Marianne Katsoyannou, and Charalambos Christodoulou. Nicosia: Cyprus Linguistics Society, pp. 1–19.
- (2025). "Ethical considerations and good practices in linguistic work on endangered languages: the case of a research programme on Cypriot Arabic". In: *Linguistics* 63.2, pp. 429–460.
- Armotistis, Spyros and Marilena Karyolemou (2023). "Contact-Induced Change in an Endangered Language: The Case of Cypriot Arabic". In: *Languages* 8.1, p. 10.

- Borg, Alexander (1985). *Cypriot Arabic: a historical and comparative investigation into the phonology and morphology of the Arabic vernacular spoken by the Maronites of Kormakiti village in the Kyrenia district of north-western Cyprus*. Stuttgart: Kommissionsverlag Steiner Wiesbaden.
- (1994). “Some evolutionary parallels and divergences in Cypriot Arabic and Maltese”. In: *Mediterranean Language Review* 8, pp. 41–67.
- (2008). “An alphabet for Cypriot Maronite Arabic”. unpublished manuscript, <https://zenodo.org/records/21086024>.
- Cahill, Michael (2014). “Non-linguistic factors in orthographies”. In: *Developing orthographies for unwritten languages*. Dallas: SIL International, pp. 9–25.
- Cahill, Michael and Elke Karan (2008). “Factors in designing effective orthographies for unwritten languages”. In: *SIL Electronic Working Papers* 2008-001.
- Christodoulou, Charalambos, Spyros Armostis, and Marilena Karyolemou (2021). “The Definiteness Marker of Cypriot Arabic: A Phonological Account”. In: *Journal of Semitic Studies* 66.2, pp. 503–521.
- Coulmas, Florian (1996). *The Blackwell encyclopedia of writing systems*. Oxford (UK) & Cambridge (Mass.): Blackwell.
- (2003). *Writing Systems: An Introduction to their Linguistic Analysis*. Cambridge: Cambridge University Press.
- Daniels, Peter and William Bright, eds. (1996). *The World’s Writing Systems*. New York.
- Fenyvesi, Anna and Marianne Bakrá-Nagy (2025). “Endangered Languages and Language Revitalization”. In: *The Handbook of Plurilingual and Intercultural Language Learning*. Ed. by Christiane Fäcke, Xuesong (Andy) Gao, and Paula Garrett-Rucks. 1st ed. Wiley, pp. 83–95.
- Fishman, Joshua A., ed. (1974). *Advances in language planning*. Berlin, Boston: De Gruyter Mouton.
- Frangeskou, Marianna and Alexander-Michael Hadjilyra (2012). *Οι Μαρωνίτες της Κύπρου [The Maronites of Cyprus]*. Nicosia: Press and Information Office, Republic of Cyprus.
- Frangiskou, Antonis G. (2000). *Το Λεξικό της Αραβικής Διαλέκτου του Κορμακίτη [The Dictionary of the Arabic Dialect of Kormakitis]*. Nicosia: Cultural Foundation of “Kormakitis” Association.
- Haarmann, Harald (2006). “Language Planning: Graphization and the Development of Writing Systems”. In: *An International Handbook of the Science of Language and Society*. Ed. by Ulrich Ammon et al. Berlin • New York: De Gruyter Mouton, pp. 2402–2420.
- Justeson, John S. (1976). “Universals of Language and Universals of Writing”. In: *Linguistic studies offered to Joseph Greenberg on the occasion of his 60th birthday*. Ed. by Alphonse Juilland, Andrew M. Devine, and Laurence D. Stephens. Vol. 1. General Linguistics. Saratoga: Amna Libri, pp. 57–94.

- Karan, Elke (2013). "The ABD of orthography testing: Integrating formal or informal testing into the orthography development process". In: *Work Papers of the Summer Institute of Linguistics, University of North Dakota Session* 53.1.
- Karyolemou, Marilena (2019). "A story at the periphery: Documenting, standardising and reviving Cypriot Arabic". In: *International Journal of the Sociology of Language* 2019.260, pp. 1–14.
- Karyolemou, Marilena and Spyros Armostis (2020). "Langue ancienne, écriture nouvelle : codification et standardisation de l'arabe de Chypre". In: *Cahiers du Centre d'Etudes Chypriotes* 50, pp. 407–429.
- Kiparsky, Paul (1982). "From cyclic Phonology to lexical phonology". In: *The structure of phonological representations*. Ed. by Harry Hulst and Norval Smith. Vol. 1. Dordrecht, The Netherlands: Foris, pp. 131–175.
- Liddicoat, Anthony J. (2005). "Corpus Planning: Syllabus and Materials Development". In: *Handbook of Research in Second Language Teaching and Learning*. Ed. by Eli Hinkel. New York: Routledge, pp. 993–1012.
- Malone, Susan E. (2004). *Manual for developing Literacy and Adult Education Programs in Minority Language Communities*. Bangkok: UNESCO.
- Meletis, Dimitrios and Christa Dürscheid (2022). *An Overview of Grapholinguistics*. Berlin, Boston: De Gruyter Mouton.
- Rogers, Henry (2005). *Writing systems: a linguistic approach*. Malden, MA: Wiley-Blackwell.
- Smalley, William Allen (1964). "How Shall I Write this Language?" In: *Orthography Studies: Articles on New Writing Systems*. Ed. by William Allen Smalley and et al. Helps for Translators. London: United Bible Societies, pp. 31–52.
- Weingarten, Rüdiger (2011). "Comparative graphematics". In: *Written Language & Literacy* 14.1, pp. 12–38.

Written Communication and the Mental Lexicon

Agglutinative Word-Form Recognition in Arabic

Joseph Dichy

Abstract. This paper is concerned with the overall structure of the mental lexicon in an agglutinative language such as Arabic. Due to that feature, as reflected by the writing of the language, a great number of nouns, adjectives and verbs do not appear as nude word-forms and include formatives on their right and/or left side. This implies that the recognition of the lexical stem of a word-form is a complex process, which is accounted for by a *word-formative grammar*. This is not without consequences for the structure of the mental lexicon, both in general and in that language.

It well-known today that the mental lexicon cannot be reduced to a mere sense/form set of features. In addition to meaning (or sense) and to phonic or graphic form, one must add a third fundamental aspect, which relates the lexical entry to its context. This is what Igor Melcuk (1982) calls the *syntactic*, which includes, e.g. gender in languages such as Latin, French or German... Verbs can be transitive or intransitive but also accept—or not—human (or quasi human) subjects or complements (Dichy, 2005).

In this paper, we consider the *syntactic* of Arabic lexical entries within the frame of the written word-form, the features of which are called *morpho-syntactic specifiers*. These include, at first, the indication of the major lexical category of the stem, due to the fact that some proclitic tool-words are related to nouns (typically, cliticised prepositions) and others to verbs (e.g. cliticised conjunctions). Complement pronouns being also cliticised, the transitivity or intransitivity of verbs must also be taken into consideration. These are basic examples. In this paper, a complete list of Arabic word-form specifiers is presented. In the DIINAR database, these have been added on a systematic basis to all the 129,000 lexical entries.

The paper illustrates the fact that morpho-syntactic specifiers are necessary in Arabic for the recognition of word-forms. They subsequently need to be added to the meaning and form of lexical entries in the mental lexicon of the language.

Joseph Dichy  0000-0002-9123-7358
Canadian University Dubai
E-mail: joseph.dichy@yahoo.fr

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 209–228. <https://doi.org/10.36824/2024-graf-dich>
ISBN: 978-2-487055-10-0, e-ISBN: 978-2-487055-11-7

1. Introduction

The structure of the mental lexicon is a crucial aspect of grapholinguistics. It is well-known that the mental lexicon needs to register both the acoustic forms of lexical entries and their written realisations, e.g.,

The English written word *read* is realised phonetically as either [ri:d] or [red] and corresponds, respectively, to either the present-tense or the preterit of that verb.

It is less known, though, that many other pieces of information are to be added for the reading process to operate, particularly in an agglutinative language such as Arabic, where:

a verb can be transitive or not, accept human subjects or human complements, or not (Dichy 1990, chapter X), and require, or not, a plural subject (Fassi-Fehri 2012), a noun can be masculine/feminine (or neutral), countable or not, collective/singulative...

Such morpho-syntactic factors directly affect the structure of the Arabic written word-form. One also needs to mention other types of information associated with lexical entries and affecting word-form recognition in the process of syntactic group or sentence understanding, such as collocations, lexical functions (Mel'cuk et al. 1984; 1995), phrasal relations, etc.

How do such relations operate? Elaborating descriptions or databases including a seemingly unlimited number of relations associated with every item of the lexicon seems at first sight extremely difficult to achieve. Natural languages encompass, in addition to limited inventories of grammatical items, an unlimited series of lexical or terminological items. The view according to which one cannot really deal with unlimited inventories can, however, be challenged.

One knows that a small number of syntactico-semantic features allow the insertion of lexical items in syntactic structures (Chomsky, 1972...). The author has shown (Dichy, 1997; 2005) how in Arabic and other languages, a *limited number* of morpho-syntactic and syntactico-semantic features restrict subject-to-verb and verb-to-complement relations in argument structures associated with verbs, e.g.,

In English, the verbs *to think* or *to believe* require a human grammatical subject (or an assimilated cognitive agent); *to read* requires a non-human grammatical object-complement¹. So do their respective Arabic equivalents, *fakkara*, *ʔiʔaqada*, and *qaraʔa*.

1. Except in the phrase *I read you*, where this verb takes another meaning, which entails considering it as another entry in a lexical database. English grammars often disregard the [±human] feature, which must, in the author's view, be added to the well-known [±animated] one. In Arabic, only the former is systematically relevant.

A phrasal verb such as *to put off* has a different meaning depending on the [$\pm$ human] feature of their complement (*to put off someone* vs. *to put off something*). In Arabic, the verb *raġiba* has two opposite meanings depending on the associated preposition, *raġiba fī* signifying “to desire” or “wish to,” and *raġiba ʿan*, to “dislike” or “detest”.

Recalling known approaches to these questions² is beyond our scope here. The essential point is that grammar-lexis relations associate unlimited—or non-finite—items, such as the entries of the lexicon of a given language, to limited—or finite—ones,³ such as grammatical categories, relations, and constraints. The fact that a finite number of relations projected on the non-finite entries of the lexicon ‘manage’ linguistic utterances is a general characteristic of natural languages, which makes it possible for either production or recognition to operate.

This general characteristic has been illustrated by the author in the recognition of Arabic word-forms. As a contribution to language theory and to grapholinguistics, we present here the set of morpho-syntactic information associated to the entries of the Arabic lexicon. These morphosyntactic pieces of information are needed for producing or recognising written word-forms.

We consider indicating how a set of information similar to that observed in Arabic operates in Biblical Hebrew or in Aramaic/Syriac word-forms in further works. The written word-forms of this family of Semitic languages can be described as *agglutinative*. In these languages, grammatical morphemes are ‘agglutinated’ in a very structured and consistent way around a single lexeme, according to what can be described as a word-formative grammar (Desclés et al. 1982; Dichy 1990). Arabic and Biblical Hebrew (as opposed to Modern Hebrew) only feature monolexical entries, a subsequent majority of which can be analysed into a Semitic consonantal root and a morphosyntactic pattern. There are a few compound entries in Arabic, but their very small number and the fact that they are only found in nominal or adjectival entries allow one to consider the phenomenon as marginal. The complex structure of the word-form, which includes the root and pattern structure, offers a good explanatory hypothesis of the mono-lexical structure of Arabic entries.

2. Let us quote, though, the semantic and cognitive approach of Desclés (1994...) and M. Abraham (1995), and, for Arabic, in a different theoretical frame, A. Fassi-Fehri (2012...).

3. Melchuk (1982, 1984); for Arabic see Dichy (1997), Dichy & Farghali (2003) and the DIINAR (DIctionnaire INformatisé de l'ARabe) data-base (Dichy, Braham, Ghazali & Hassoun, 2002).

1.1. Structure of the Agglutinative Written Word-Form in Arabic

In the agglutinative graphic word-form structure of Arabic, one encounters a number of grammatical lexemes and a single lexical nucleus, according to the example below:

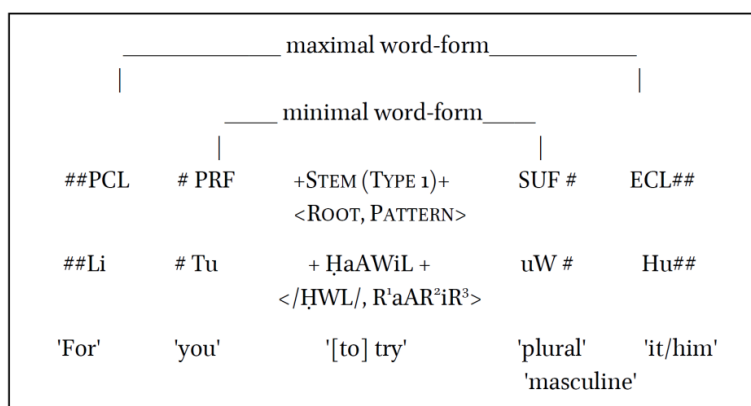


FIG. 1. Scheme of the graphic Arabic word-form in Arabic

Word-Form Codes Used in the Above Table and Below

1. Word formatives

STEMS or *nuclei* are either analysable into ROOTS and PATTERNS (type 1), as in the example above, or not (type 2), as with loan-words or a great number of proper names.

PCL = proclitic(s), PRF = prefix, SUF = suffixe(s), ECL = enclitic(s);

2. Word-formative boundaries

+ = strong word-formative boundary, or lexical nucleus boundary,

= weak word-formative boundary or clitic boundary, occurring after a PCL or before an ECL.

3. ROOT and PATTERN codes

R¹, R², R³, or R1, R2, R3, respectively 1st, 2nd and 3rd radical consonant in 3-consonantal roots. R⁴ or R4 is added in 4-consonantal roots. A PATTERN can be represented (a) either with these numbered radical consonants, to which other elements, such as short or long vowels, are added or (b) with the traditional Arabic grammar code, using letters /f/, /ʕ/ and /l/, and repeating /l/ in 4-consonant ROOTS, e.g., in 3-consonant ROOTS: *makātib*, “offices”: (a) maR1āR2iR3 and (b) mafāʕil and in 4-consonant ones: *ʿarqala*, “to hinder, impede” or “obstruct”: (a) R1aR2R3aR4a and (b) faʕlala.

4. Transcription codes only used in the above table

Capital letters transcribe Arabic graphemes written in the body of word-forms, while *lower-case letters* transcribe diacritical short vowel signs.

The frame presented in Fig. 1, primitively introduced by David Cohen (1961), has been taken up and developed in (Desclés et al. 1982; Dichy 1990) and so forth.

For the lexical nucleus to be correctly inserted into word-forms, it must be associated with *morpho-syntactic specifiers* (introduced in Dichy 1990 and 1997), which account for relations between the nucleus and the other formatives situated right and left of it in the graphic word-form. They need to be associated with every verb, noun and adjective of the language in the speaker's mental lexicon, or in a lexical database, such as the one designed and achieved in the second part of the 1990's for the DIINAR database (Dictionnaire INformatisé de l'ARABE)/معالي (معجم العربية الآلي)—between researchers located in Lyon and a high-level Arab research centre in Tunis⁴. These specifiers are limited in number: they account for relations between lexical nuclei and grammatical lexemes included in graphic word-forms. The fact that grammatical lexemes belong to closed lists accounts for the limited number of specifiers, which need in turn to be added to every lexical entry.

Psycholinguistic research on the recognition of written Hebrew (Frost et al., 1997, Frost et al., 2000) and written Arabic (Grainger, Dichy et al., 2003) has shown that the reading process in these languages is primarily based on contour recognition. This is highly in accordance with the formal analysis introduced here. It shows that the Semitic word-form reading process is based on closed-list information associated with every lexical item included in the mental lexicon of readers. These information items are called morpho-lexical specifiers.

2. Arabic Word Formatives and Associated Specifiers

2.1. Word-Form Nuclei Categories

In the DIINAR database, we have considered the category of verbs on one side and that of nouns and adjectives on the other, following the medieval Arabic grammatical tradition. Although nouns and adjectives need to be distinguished syntactically and semantically, we have found

4. *Institut Régional des Sciences de l'Informatique et des Communications* ("The Regional Institute of Computer Science and Tele-Communications"), Tunis, Profs. Salem Ghazeli and Abdelfattah Braham.

that the nouns/adjectives category remains sufficient for the recognition of word-forms. Adverbs, such as *ʿayḍan*, “also, too” or *laylan*, “nightly, by night” are included in the noun category.

2.1.1. *Lexical Nuclei*

Main Features

The main feature of the lexical nucleus is that it is either:

- unrelated to the root and pattern derivation system, such as nouns pertaining to the following categories:
 - proper names (generally borrowed, including anciently), e.g., *Butrus* (from the Latin *Petrus*, *Peter* in English), *Maryam*, *ʾIsmāʿīl*, *Yū-nescū* (UNESCO), *Sūriya*, “Syria,” or
 - loan-words, eg.: *banāna*, “banana,” *makīna*, “machine,” *mūbāyl* (“mobile phone”),
- representable in terms of root and pattern, such as:
 - nouns/adjectives, e.g., *karīm*, “noble and generous,” ROOT /K-R-M/, PATTERN *faʿīl*, or R1aR2iR3, where R1 = /K/, R2 = /R/, R3 = /M/,
 - verbs, e.g., *rasīla*, “to be long and flowing” [hair], which can be analysed into ROOT /R-S-L/, and PATTERN *faʿīla* or R1aR2iR3a.

Number of Radicals in ROOTS

I have shown (Dichy, 1990), that the Arabic language only features 3-consonant and 4-consonant roots, the definition criterion in Semitic languages being that roots must be subject to semitic-morphological derivation with simple and augmented patterns, e.g.,

- the 5-consonant word *barnāmaj*, “roster,” or “program” has a 4-consonant plural, *barāmij*, featuring an ad-hoc 4-consonants ROOT /B-R-M-J/, which allowed, in modern times, the verb *barmaja*, “to program”;
- *safarjal* “quince” has for a plural *safārij* (ad-hoc 4-c. ROOT /S-F-R-J/);
- the 2-consonant word *yad* “hand” has 3-consonants plurals: *ʿayādin* and *ʿaydin* (ad-hoc 3-consonants ROOT /Y-D-Y/); medieval dictionaries also mention *ʿayyada* and *yadda*, “giving a hand, helping and supporting” (ad-hoc ROOTS /ʾ-Y-D/ and /Y-D-D/).

Also note that *all* verbs and verb-derived nouns and adjectives pertain to the Semitic root-and-pattern frame, which, on the other hand, only includes a subset of nouns and adjectives (Dichy, 1990).

Consistency of Nucleus Representation

In this representation, consonants belonging to the pattern and added to the root are considered as part of the lexical nucleus, whether they appear between the ROOT consonants, e.g.,

- *wāw* inserted before the last ROOT consonant, as in *durūs*, “lessons,” or whether they appear at the beginning or the end of the PATTERN, e.g.,
- /a&/⁵ added at the end of *kabāma&*, “dullness” or “lassitude, weakness”; *ʾistiḳāma&*, “honesty”;
- *madraj*, “course,” “route,” “path,” “road” ...⁶

This representation has the advantage of avoiding splitting the word-form nucleus, and of featuring it, for the sake of consistency, as a full entry of the mental lexicon or of a lexical database. Morpho-syntactic or syntactico-semantic specifiers are associated with that entry as a whole.

2.1.2. *Tool-Words or Void Nuclei*

Tool-words can also appear in the position of a word-form nucleus, e.g.,

- *wa#inna#bum*, where “#” is for clitic boundary, *inna* is a conjunction, and *#bum* is the plural masculine complement pronoun.

Some word-forms do not include a nucleus, e.g.,

- *La#bum* “for them,” *bi#bi* “by him or it”.

When the ECL pronoun is syntactically marked as a direct-case pronoun, an epenthetic nucleus appears, e.g.,

- *ʾanā wa#ʾiyyā#ka*, “me and you” [as one]; Koran I,5: *ʾiyyā#ka naʿbud+u*, “Thee [only] we worship”.⁷ Another example is the use of *miṭl*, “like” or “likeness” in the phrase-word *ka#miṭl+i#bi* “similar to” (Koran 42/11: *laysa kamiṭlibi šayʾun*, “There is naught to the likeness of Him”).

5. By convention (after D. Cohen 1961), the *tāʾ marbūʿa* is represented as *a&*. There are two distinct representations (Desclés et al., 1982, Dichy 1990):

- without a suffix boundary as in the example above, where it belongs to the pattern, or
- with a strong boundary symbol “+,” i.e., *+a&* when it is a feminine gender suffix, e.g., *jamil+a&*, “pretty,” feminine, or a singulative SUF (*ʾism al-wāḥid*), as in: *naml+a&*, “ant,” plur. *naml+āt*.

6. For the analysis of these two pattern ‘infixes’ (*m+* and *+a&*) identified as 1-consonantal roots, see A. Roman (2011).

7. Current Arabic grammars present the sequence *<ʾiyyā+complement pronoun>* as an “isolated direct case pronouns” (*ḍamāʾir munfaṣila manṣūba*), which is doubtful, since isolated pronouns are always of the subject case.

In this paper, we do not present the tables of compatibility between tool-words and proclitics and/or enclitics (ECL), due to the fact that every tool-word category has its own compatibility rules. In fact, these tables, which I devised for the DIINAR database and associated analysers, allow the building of a generated lexicon for Arabic natural language processing (NLP), which only includes extant items.

2.2. Prefixes (PRFs)

Prefixes are, after D. Cohen (1961/70), strictly verbal and only appear in the conjugation of the imperfect or the imperative. These are, for the imperfect:

- *ʾa+* or *ʾu+*. This PRF is for the first-person singular;
- *na+* or *nu+*, first-person plural;
- *ta+* or *tu+*, which is for the second person masculine/feminine, as well as for the third person feminine;
- *ya+* or *yu+*, which is for the third person masculine/feminine.

These prefixes are in fact personal pronouns indicating the first, second, or third person (masc. or fem.). Only *ʾa+* and *na+* indicate a difference in number (singular vs. plural). The selection of the vowel *a* or *u* above depends on the morpho-syntactic pattern of the verb. Vowel *u* appears with patterns *faʿala*, *ʾafʿala* and *fāʿala* or with the passive voice.

For the imperative:

- *ʾu+* or *ʾi+* is an epenthetic *hamza* appearing at the beginning of the imperative verbs pertaining to the simple pattern and in some augmented patterns, with a variation of the vowel according to a well-known grammar rule (due to apophony).

2.3. Suffixes (SUFs)

SUFs are divided into nominal/adjectival and verbal ones. In the same way as prefixes, their presence in the word-form is morphologically necessary: all lexical entries have suffixes, the minimal suffixes in nouns being the case ending, and in verbs, gender and number desinences. This is formally represented by the “+” strong boundary symbol.

2.3.1. Verbal SUFs

These suffixes are related to conjugation paradigms, which have been modeled in Ammar, and Dichy (1999/2008), in accordance with the representation devised for the conjugation of verbs in various languages

by the French grammarian Bescherelle. His basic idea is that a small number of verb models (ranging from 88 to over 200) can allow the conjugation by analogy of all the verbs of a language, provided the rules of the language are modeled in such a way as to produce a strict formal classification of verbs in relation to conjugation. (In today's terms, this is described as a taxonomy, i.e., a formal classification where a given item appears in one slot and one only.) Regarding the Arabic language, I have designed a 2-dimensional table featuring horizontally the simple and augmented verb patterns, and vertically the 33 types of 3-consonantal roots, according to the number of the transformation causes they include, such as a radical consonant *hamza*, *waw* or *ya*, two identical consonants in second and third position ($R_2 = R_3$), and two or more of these features within a given root.⁸ Another table accounts for 4-consonantal roots.

The conjugation of Arabic, complex as it may seem with certain verbs, is regular (with a very small number of exceptions). There are, subsequently, two ways of answering any question related to it:

- by applying the rules, which requires a certain level of calculation, and is likely to occur as a remediation strategy, when the speaker encounters a difficulty;
- by analogy, through imitating the conjugation of a similar verb. This is the idea brought forth by Louis-Nicolas Bescherelle in the 19th century, which is applied in the Bescherelle series of conjugation reference books for many languages, published in France today by Hatier. It includes Ammar & Dichy's *Bescherelle des verbes arabes* (1999/2008). The 125 verb models identified in this work are numbered, and the volume includes a lexicon of verbs, each entry of which is associated with the number of the model according to which it can be conjugated.

Human memory may well operate by analogy, with the possibility for speakers to resort to calculation when they encounter a difficult point. Whatever it is, Ammar & Dichy's analogical models (1999/2008 and 2021)⁹ perfectly fit conjugation systems assisted by a database.

8. See Dichy 1993; Ammar & Dichy 2008, p. 34-36; 2021, chapter 1. By contrast, other conjugation presentations include hundreds of models, while not covering the whole conjugation of the language (see Bayyūmī et al. 1989; Osman 1995; Reig 1983; Sheidlin 1978).

9. The former volume (1999/2008) includes a lexicon of 10,000 verbs, and a short grammar; the latter (2021) has a reference lexicon of just under 20,000 verbs and includes many grammatical remarks in relation with the verb models, and a detailed theoretical presentation of the analysis that led to a complete coverage of Arabic conjugation.

Conjugation Morphosyntactic Specifiers

In the DIINAR database and its analysers¹⁰, the numbers associated with every verb and referring to the 125 analogical models of Ammar & Dichy (1999/2008) are considered as morphosyntactic specifiers related to conjugation. A pattern-matching generation program based on these models has been designed and implemented by R. Zaafrani (2002).

Other specifiers related to conjugation are concerned with the passive voice, the imperative, or the absence of active and/or passive participles:

- Some transitive verbs can only be conjugated in the passive voice at the third person. This occurs when the verb is transitive to non-human complements, e.g.,
 - *waba'a* "to prepare" [luggage], *hayya* "to live" [something, an experience].

In such verbs, the non-human complement cannot "speak" (1st person) or "be addressed" (2nd person) in the passive voice.

- Some intransitive verbs, or prepositional transitive verbs, can only be conjugated in the passive voice at the third person masculine singular, e.g.,
 - *furiḥa* "to be glad," "to rejoice"; *wali'a* "to be madly in love (with)," "to be passionately fond (of)," followed in Arabic by the preposition *bi*#.

In some verbs, in the passive voice, the first and the second person are strictly excluded for semantic reasons, and the 3rd person expresses 'impersonal passive' phrasing, meaning, word-for-word:

- *furiḥa*, "One has rejoiced," i.e., "People rejoiced"; *fatātun kaṭīran mā wuli'a biḥā*, "a maiden many-a-time [men] fell in love with"...
- Some verbs are both intransitive and non-agentive, which excludes the passive voice and the imperative, e.g.,
 - *saruwa* "to be of noble descent"; *yamuna* "to be blessed by luck"; *ra'ufa* "to be of merciful nature"; *qabuḥa* "to be ugly".
- In addition to the preceding constraint, some strictly non-agentive verbs do not have an active and/or passive participle, e.g.,
 - *ṣaḡura* "to be small," for which one can neither say *ṣaḡīr* in this meaning of the verb, nor **maṣḡūr*. Instead, one uses the analogous adjective (*ṣifa muṣabbaba*) *ṣaḡīr*.
 - Other analogous adjectives are related to some of the above examples, e.g., *ra'ūf* "merciful (by nature)" for *ra'ufa* or *qabīḥ* "ugly" for *qabuḥa*.

10. See the doctoral dissertations co-supervised by the author of Ghenima (1998), Zaafrani (2002), Ouersighni (2002), Raheel (2010).

2.3.2. Nominal/Adjectival SUFFs

Nominal/adjectival suffixes are of five general types:

- gender and number suffixes, such as:
 - *+a&* (feminine singular), *+āt* (feminine plural), *+ā'* (feminine singular in some adjectives) *+ūna* (masculine plural), *+ū* (masculine plural, elided *na*)...;
- case-ending suffixes, such as:
 - *+un* (subject case, indefinite), *+u* (subject case, indefinite, *n* elided), *+an* (direct case, indefinite, written with *alif*), *+a* (direct case, indefinite, written with *alif*, *n* elided), *+in* (indirect case, indefinite), *+i* (indirect case, indefinite, *n* elided), etc.;
- compound gender-number and case-ending suffixes, such as:
 - *+ūna* (masc. pl., subject case) and *+īna* (masc. pl., direct and indirect cases); *+āni* (dual number, subject case) and *+ayni* (dual number, direct and indirect cases); *+āt+un* (fem. pl., subject case) and *+āt+in* (fem. pl., direct and indirect cases), etc.
- the suffix *+iyy* of the relative noun-adjective (*yā'* 'al-nisba');
- the compound suffixes *+iyya* and *+iyyāt* are used in modern Arabic for abstract notions, e.g.,
 - *silmi+iyya*, "pacifism"; *lisān+iyyāt*, "linguistics".

Morphosyntactic Specifiers Related to Nominal/Adjectival SUFFs

These suffixes are subject to lexical or grammatical constraints, and require specifiers associated with lexical entries:

- Indication of gender and number suffixes must be given systematically, e.g.,
 - *mudarris*, "teacher," masc. plur., *mudarris+ūna/mudarris+īna*; the feminine is *mudarrisa&un*, fem. plur., *mudarris+āt+un/mudarrisāt+in*.
- A great number of plural forms are not built with suffixes but require a change in the pattern of the noun or adjective. These are traditionally referred to as 'broken plural' forms. When this occurs, it requires that, in the database, plural and singular forms be registered together and that the relation between the singular and the broken plural be 'arrowed', e.g.,
 - *kurs+iyy*, "chair," broken plural, *karās+in*; *safīr*, "ambassador," plural *sufarā'*; adjective *karīm*, "noble and generous," plural forms *kuramā'* and *kirām*...
- 'Arrowing' also applies to the feminine and plural of many adjectives, such as:
 - 'azraq "blue," fem. sing. *zaraqā'*, pl. *zurq*.

- In some cases, the plural is in modern Arabic built with either a suffix or a change in the pattern (broken plural), e.g.,
 - *jamīl* “beautiful, pretty,” masc. pl. for humans *jamīl+ūna*, fem. *jamīl+āt+un*, shared broken plural *jumalāʾ*.

This requires indications of suffixes, broken plurals, or feminine forms featuring a change in pattern (as with *ʿazraq* above) that cover all cases. These relations are, of course, limited in number.

- Case endings vary with a subset of nouns and adjectives, known as diptotes (*mamnūʿ mina al-ṣarf*), e.g.,

- *ʿaswad*, “black,” fem. *sawdaʾ*, do not take all the case endings.

This requires a specific morphosyntactic specifier, due to the fact that not all diptotes are predictable, e.g.,

- the proper noun *Nibād* is a diptote when it is feminine, and not when it is masculine; *ʿašyāʾ* “things”, is lexically coded as a diptote.

- The relative suffix *+iyy* does not appear with all nouns or adjectives, due to usage. This means that its occurrence with a given entry needs to be recorded, either as an additional lexical entry or with an associated specifier. E.g.:

- *bayrūt*, “Beirut” allows *bayrūt+iyy*, “related to, or inhabitant of Beirut”.

- On the other hand, *Dubay*, “Dubai” or *Jibūṭī*, “Djibuti,” which end with a *y*, do not accept the suffix *+iyy*. This entails the use of suppletive forms, such as *ʿabl Dubay* or *jibūṭī*, “inhabitants of Dubai or Djibouti”; the same can be said, for other reasons of *Abū Dabī* or *Rās al-ḥayma*...

- The suffix *+iyy* may entail a modification of the noun it is suffixed to, e.g.,

- *qabīla* “tribe” becomes *qabal+*, when *+iyy* is suffixed: *qabal+iyy*, “tribal”.

- In contrast, *ṭabīʿa* “nature” is not affected in its structure by this suffix: *ṭabīʿ+iyy*, “natural”.

- Some proper nouns feature a more radical change in their structure, e.g.,

- *al-quds*, the Arabic name for “Jerusalem,” has a relative noun-adjective *maqdis+iyy* (built after the ancient Islamic name of the city *Bayt el-Maqdis*). The adjective *quds+iyy* being already taken in the lexicon, with the meaning of “sacred,” in contemporary Arabic, one says *ʿabl* or *sukān al-quds*, “people or inhabitants of Jerusalem [Arabic name]”.

- The city of Casablanca is called in Arabic *al-dar al-bayḍāʾ*, the relative noun-adjective being *bayḍāw+iyy*.

In the same way as with broken plurals, an ‘arrowing’ relation must, when applicable, link the stem of the noun with and without the suffix *+iyy*.

- When the noun ends with /+a&/, the introduction of +iyy entails the deletion of that marker, e.g.,
 - *jāmi*‘a&, “university” becomes *jāmi*‘+iyy, “academic,” “related to a university”.

This allows the <noun+iyy> compound, shaped like a nūn to have a feminine form <noun+iyy+a&>, avoiding a ‘doom loop’ such as:

*<noun+a&+iyy+a&+iyya...>.

2.3.3. *Proclitics (PCLs)*

PCLs are of various types. Basically, they are 1-consonant tool-words (the article *al*# being a specific case, featuring an /l/ consonant preceded by an epenthetic hamza).

The main difference between prefixes and suffixes, which are included in minimal words (see Fig. 1), is that PCLs are added to them without being morphologically necessary to the structure of the word-form. This is formally represented by the # weak boundary symbol.

PCLs are divided into three general categories, the first of which is common to all major category nuclei, the second and the third being respectively related to nouns and verbs.

PCLs are submitted to a strict order of occurrence in the word-form, as well as to a number of grammar rules. Both fall under the word formative grammar (Desclés & al, 1982; Dichy 1990, chap. X, §3.4).

Two features are related to the readability of word-forms and facilitate word-form recognition:

1. PCLs belong to a limited set of consonants: /ʔ/, /w/, /f/, /k/, /b/, /l/, /ʔā/, /s/, /l/, to which one must add the article /ʔal/.
2. They appear according to the following strict order:
 - Position 1: /ʔa#/, interrogative
 - Position 2: /wa#/, /fa#/, coordinating conjunctions;
 - Position 3: /la#/ corroborative conjunction (*lām al-tawkid*)
 - Position 4: /ka#/, comparative particle
 - Position 5: /bi#/, /li#/, prepositions
 - Position 6 and 7: /ʔal#/, definite article
 - Position 7: /li#/, energetic particle, which loses its vowel /i/ when combined with the coordinating conjunctions *wa#* or *fa#* and produces: /wa#l#/, /fa#l#/; /sa#/, future marker; /wa#/, *wāw rubba*; /ʔa#/, /ʔā#/ vocative particles.

The above list includes PCLs that one does not expect to occur in contemporary Arabic. It is nevertheless essential to retain them, the system needing to be sturdy enough to analyse both modern and classical corpora.

2.3.4. *PCLs Shared by All Major Categories*

Four PCLs appear at the beginning of a word-form (positions 1-3) according to the above-mentioned strict order.

The first shared category of PCLs is the interrogative tool-word *hamza*, which always comes first when occurring in a word-form.

The second category in the word form order of PCLs is that of the 1-consonant coordinative conjunctions, i.e., *wa#* “and,” *fa#* “and then,” and of the corroborative conjunction *la#*, which can only be translated into English through the addition of an adverb such as “surely,” “indeed,” or “certainly”.

2.3.5. *Verbal PCLs*

In position 7, one finds 2 verbal PCLs,

- the energetic particle *li#* usually followed by a jussive imperfective: e.g., *li#na+ktub*, “let us write”. One must also add the compound PCLs /*wa#l#*/ and /*fa#l#*/
- the future marker *sa#*, e.g., *sa#na+ktub*, “we shall write”.

2.3.6. *Nominal PCLs*

These include:

- the comparative tool-word *ka#* “as, like”;
- the 1-consonant prepositions *bi#* “with, along, by...,” *li#* “to, for...”
- the article *al#*;
- 1-consonant tool-words, mainly found in ancient poetry, *wa#*, *wāw rubba*, meaning “many-a,” e.g.,
 - *wa#layl+in ka#mawj+I l#baḥr+i...*, “Many-a night alike sea-waves...,”
- and the vocative particles /^ʔa/, /^ʔā/, e.g.,
 - ^ʔa#*fātima mabl+an...*, “O Fatima, slowly...” (examples from Imruʔu al-Qays’s Ode).

2.3.7. *Category Selection*

PCLs mentioned in the two paragraphs above select the major category (noun, verb) they can appear with. Verbal PCLs also select the conjugation paradigm (imperfective, jussive, and imperative).

These selections pertain to the word-formative grammar.

2.4. Enclitics (ECLs)

These are the attached complement pronouns. Verbal ECLs can only appear with transitive verbs, and highlight the necessity of adding a morpho-syntactic specifier indicating that:

- either the verb is only transitive to non-human complements, e.g., *qara'a*, “to read” seen above. This excludes pronouns referring to humans, such as the masculine plural *#bum*;
- or that the transitive verb admits human complements, in which case all ECLs can appear.

2.5. Double Enclitics

Some verbs admit two direct complements, e.g.,

- *ʔaʕtā*, “to give” which admits *ʔaʕtayt+u#ka#bu*, “I gave it to you,” or the performative marriage formula: *zawwaj+tu#ka#bā*, “I give her in marriage to you”.

Sequences of double enclitics are restricted for euphonic reasons: one cannot find, e.g., **ʔaʕtayt+u#ba#ba*. These restrictions are also included in the word-formatives grammar.

3. Lexicalized formatives (Dichy, 1997)

Some formatives attached to nominal nuclei present what can be described as a cluster structure usually occurring in proper nouns, e.g., *al#Tayyib* (masculine example); *al#Zabrāʔ* (feminine example). Another example is that of *al#Babr+ayn*, the name of Bahrein, originally meaning “the two seas”.

4. Conclusion

Language specifiers restrict the number of items in the generated lexicons of the language. Let us take, as an example, the DIINAR database. The number of *extant* entries or nuclei is 131,087, which generates 7,764,937 *attestable* inflected minimal word-forms, instead of between 45 and 70 million virtual ones with a blind generation system¹¹ (Dichy et al., 2004; 2005). The breakdown is the following:¹²

11. I do not present here a critical view of ancient and even current research in that domain.

12. Column 3 stands for “GENERATED LEXICA of minimal words (inflected forms)”.

TABLE 1. Number of entries in the generated lexica of DIINAR, for minimal word-forms

	Stems or nuclei	GEN. LEX.
Verbs	19,457	3,050,715
Deverbals = immediate verb-derived nominal forms, such as: مصادر، أسماء الفاعل والمفعول، صفات مشبهة، إلخ	70,702	2,909,772
Nouns and adjectives (including broken plurals)	39,099	1,781,316
Proper names (limited prototype)	1,384	11,403
Tool-words (provisional figure)	445	11,731
Total	131,087	7,764,937

In this database, the morpho-syntactic specifiers accounting for the relations between the nucleus of the word-form designed by me (Dichy, 1990; 1997) and listed in the present paper, have been added to each of the entries. These specifiers restrict the number of minimal words (i.e., of entries with their prefix and/or suffixes) generated from the database to 7,764,937. All forms included in the generated lexica of DIINAR are *extant* words. On the contrary, an unconstrained combination of all word formatives with the total number of nuclei in the figure above would lead to a generated lexicon of 'virtual' maximal word-forms amounting to 57,642,863,616! Limiting the figures to minimal word-forms, the combination would still yield:

$$8 \text{ prefixes} \times 67 \text{ suffixes} \times 131\,087 \text{ stems} = 70,262,632 \text{ 'virtual' forms}^{13}$$

This shows the efficiency of the representation implemented in the DIINAR dB in restricting the analyses to real Arabic data.

It is essential to note that restricting general lexica is crucial in the natural language processing of languages (NLP), because the related software needs it to:

- operate correctly in recognition, i.e., to recognize extant word-form, rejecting others, even if they can be generated virtually;
- produce correct word-forms included in the mental lexicon, which implies that the morphosyntactic specifiers presented in this paper should be considered and integrated.

13. The consistency of these figures can be easily explained based on the analysis of the word-form and of its constituents (Dichy, 1990). Complementary word-analysis software was developed through 5 doctoral dissertations supervised or co-supervised by me between 1998 and 2010 (see note 10).

The above description shows the consistency of the constraints on the structure of word-forms in Arabic and partly explains the process of readability in unvowelled writing and the related mental lexicon in that family of Semitic languages.

References

- Abraham, Maryvonne (1995). "Contribution méthodologique à la constitution d'un dictionnaire informatique des verbes : analyse sémantico-cognitive des verbes de mouvement et d'activité". PhD thesis. Paris: EHESS.
- Ammar, Sam and Joseph Dichy (1999). *Bescherelle des verbes arabe*. Édition revue et actualisée 2008. Paris: Hatier.
- (2021). *20001 verbes arabes. Dictionnaire de conjugaison, précédé des principes de classification et de modélisation de conjugaison*. Beyrouth: Librairie du Liban.
- Bayyūmī, Ḥasan, Kalfat Khalīl, and Aḥmad al-Shāfiʿī (1989). بيومي، حسن، خليل كلفت، أحمد الشافعي. معجم تصريف الأفعال العربية [Dictionary of Arabic Verb Conjugations]. دار إلياس العصرية: القاهرة [Cairo].
- Braham, Abdelfattah and Salem Ghazali (1998). "ابراهيم، عبدالفتاح وسالم غزالي". قاعدة البيانات المعجمية العربية أو مشروع معجم العربية الآلي (معالي—DIINAR)، حصيلة وآفاق [Arabic Lexical Database or Automated Arabic Dictionary Project]. In: المجلة العربية للعلوم [Arab Journal of Science] 32, pp. 14–23.
- Chomsky, N. (1972). *Studies on Semantics in Generative Grammar*. Trad. fr. *Questions de sémantique*, Paris, Le Seuil, 1975. La Haye: Mouton.
- Cohen, David (1970). "Essai d'une analyse automatique de l'arabe". In: *Études de linguistique sémitique et arabe*. Originally published in *T.A. informations*, 1961. Paris: Mouton.
- Deroche, François (2004). *Le livre manuscrit arabe. Préludes à une histoire*. Paris: Bnf.
- Desclés, Jean-Pierre (1994). "Relations casuelles et schèmes sémantico-cognitifs". In: *Langages* 113. *Les relations actanciels. Sémantique, syntaxe, morphologie*, pp. 113–125.
- Desclés, Jean-Pierre et al. (1982). *Conception d'un synthétiseur et d'un analyseur morphologiques de l'arabe, en vue d'une utilisation en Enseignement assisté par Ordinateur*. Tech. rep. Rapport rédigé sous la direction de J.-P. Desclés pour le Ministère français des Affaires étrangères (sous-direction de la Politique linguistique).
- Dichy, Joseph (1990). "L'écriture dans la représentation de la langue : la lettre et le mot en arabe". 2 vol. Doctorat d'État. Université de Lyon.

- Dichy, Joseph (1993). "Knowledge-system simulation and the computer-aided learning of Arabic verb-form synthesis and analysis". In: *Processing Arabic Report* 6/7, pp. 67–84, 92–95.
- (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. URL: <http://www.erudit.org/revue/meta/1997/v42/n2/002564ar.pdf>.
- (2005). "Spécificateurs engendrés par les traits [$\pm$ ANIME], [$\pm$ HUMAIN], [$\pm$ CONCRET] et structures d'arguments en arabe et en français". In: *De la mesure dans les termes. Hommage à Philippe Thoiron*. Ed. by Henri Béjoint and François Maniez. Presses Universitaires de Lyon, pp. 152–181.
- (2022). "Semitic Writings and Short Vowels: Alternative Hypotheses in a Renewed View of the Analytics of Writing". In: *Proceedings of Grapholinguistics in the 21st Century, 2022*. Ed. by Yannis Haralambous. Vol. 9. Grapholinguistics and Its Applications. Brest: Fluxus Editions, pp. 377–390. DOI: <https://doi.org/10.36824/2022-graf-dichy>.
- Dichy, Joseph and Ali Ferghali (2003). "Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical database centred on Arabic be built?" In: *Proceedings of the Workshop on Machine Translation for Semitic Languages: Issues and Approaches*. IXth Machine Translation Summit, New Orleans, 23–27 September 2003, pp. 1–8.
- Dichy, Joseph and Mohamed Hassoun (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*. Ed. by Abdelfattah Braham. La Manouba-Tunis: Université de La Manouba, pp. 45–56.
- Fassi Fehri, Abdelkader (2012). *Key Features and Parameters in Arabic Grammar*. The Netherlands/Philadelphia: John Benjamins.
- 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. 751–765.
- 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.
- Ghenima, Malek (1998). "Analyse morpho-syntaxique en vue de la voyelation assistée par ordinateur des textes écrits en arabe". PhD thesis. ENSSIB/Université Lyon.

- 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. Vol. 22, pp. 77–86.
- al-Hāḡib, Ibn (1975). "ابن الحاجب. الشافية في علمي التصريف والخط" [al-Shāfiya on the Two Sciences of Morphology and Orthography]. In: رضي الدين [Commentary on al-Shāfiya]. Ed. by Nūr al-Ḥasan et al. 4 vols. بيروت [Beirut]: [House of Scientific Books].
- Jinnī, Ibn (1985). "ابن جني. سر صناعة الإعراب، تحقيق حسن هندراوي" [The Secret of the Craft of Desinential Inflection]. Ed. by Hassan Hindāwī. 2 vols. دمشق [Damascus]: دار القلم [House of the Qalam].
- Mel'čuk, I., A. Clas, and A. Polguère (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.
- Mel'čuk, I. and et al. (1984). *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I*. Presses de l'Université de Montréal.
- Mel'čuk, Igor (1982). *Towards a Language of Linguistics, A System of Formal Notions for Theoretical Morphology*. München: Wilhelm Fink Verlag.
- Osman, Nabil (1995). *Konjugation Lexikon Arabische Verben*. München: Max Hueber Verlag.
- Ouersighni, Riadh (2001). "A major offshoot of the DIINAR-MBC project: AraParse, a morpho-syntactic analyzer of unvowelled Arabic texts". In: *ACL 39th Annual Meeting, Workshop on Arabic Language Processing: Status and Prospect*. Toulouse, pp. 66–72.
- (2002). "La conception et la réalisation d'un système d'analyse morpho-syntaxique robuste pour l'arabe : utilisation pour la détection et le diagnostic des fautes d'accord". PhD thesis. ENSSIB/Université de Lyon.
- Raheel, Saeed (2010). "Apprentissage artificiel et fouille de données multilingues : le cas des documents arabes". Doctorat en Sciences de l'Information et de la Communication. Université de Lyon.
- Raheel, Saeed and Joseph Dichy (2010). "An Empirical Study on the Feature's Type Effect on the Automatic Classification of Arabic Documents". In: *Computational linguistics and intelligent text processing, Proceedings of the 11th international conference, CICling 2010*. Ed. by Alexander Gelbukh. Iași, Romania, March 2010. Berlin/Heidelberg/New York: Springer-Verlag, pp. 673–686.
- Reig, Daniel (1983). *La conjugaison de la langue arabe*. Paris: Maisonneuve et Larose.
- Roman, André (2011). *Grammaire systématique de la langue arabe*. Paris: L'Harmattan.

-
- Sheidlin, Raymond (1978). *201 Arabic verbs fully conjugated in all the forms*. New York: Barron's Educational Series.
- Zaafrani, Riadh (2002). "Développement d'un environnement interactif d'apprentissage avec ordinateur de l'arabe langue étrangère". PhD thesis. ENSSIB/Université de Lyon.

Why the Abjad Suits the Arabs

Guy Boursier

Abstract. The initiators of the Wikipedia Egyptian Arabic edition strove to differentiate it from the Standard Arabic edition, notably via the script chosen. Some proposed a script adapted from the Latin alphabet, thus repeating previous attempts to systematically transcribe vowels, which the traditional Arabic abjad does not. Oddly enough, most Egyptian Arabic Wikipedians, despite their desire to cultivate their difference from Standard Arabic, chose the Arabic script used as an abjad for writing their encyclopedia's articles. Besides the weight of habit, the diglossic status of Arabic may explain this choice. Yet, they have managed to adapt the script to give it a specific "Egyptian" flavor. This article is inspired from some of the results of the research presented in the author's master's thesis, entitled *Écrire l'arabe égyptien, le cas d'un Wikipédia 'dialectal'*, defended at UFR ALL-Nancy, Université de Lorraine, France, in June 2025.

1. Standard Arabic and Egyptian Arabic

In his seminal 1959 article entitled "Diglossia," Charles A. Ferguson takes Egyptian Arabic as an example of a "low" variety in this type of situation where "two varieties of a language exist side by side through the community, with each having a definite role to play" (Ferguson, 1959). Egyptians, therefore, speak what is often described as a dialect of Arabic in their daily lives, while writing is primarily in the "high" variety, in this case "Classical Arabic," or its modern version, thence referred to as "Standard Arabic".

It is this variety that children learn to write in school, in which masterpieces are composed, such as the works of Naguib Mahfouz, laureate of the Nobel Prize for Literature in 1988. Most instances of written Egyptian Arabic have traditionally been confined to marginal items such as intimate correspondence, captions on cartoons and comic literature¹. However, recently there have been increasing attempts to cross

Guy Boursier  0009-0007-0445-376X
Aix Marseille Univ, CNRS, IREMAM, Aix-en-Provence, France
E-mail: guy.boursier@etu.univ-amu.fr

1. For an exhaustive study of diglossia in Arabic, see Diem (1974).

the border, with novels written entirely in Egyptian Arabic² or even a work translated into this variety from Classical Arabic³.

In 2008, an edition of Wikipedia in Egyptian Arabic was created (thence shortened to Wikipedia maṣri, from an adjective meaning “Egyptian,” used in the website’s language menu), parallel to the 2004 Arabic edition (Wikipedia al-‘arabiyya). This constituted a further step towards the dialect’s legitimation as a language in its own right, capable of dealing with all the fields of human knowledge to be found in an encyclopedia. However, as with any emergence of a written corpus, many questions needed to be resolved.

2. How to Write Egyptian Arabic?

A major question is: how to write? Of course, the traditional way of writing Arabic is the Arabic script used mainly as an abjad, a system in which mostly consonants are represented, leaving vowels to be deduced by the reader, as seen in Table 1. For sure, vowels can be indicated with diacritic marks, but these are rarely used. The Arabic script can also be used so as to write consonants and vowels phonographically, essentially in loan words such as *dimuqrāṭiyya*—ديمقراطية (democracy), but this is marginal in Standard and Egyptian Arabic. One of the *raisons d’être* of abjad is to make the writer gain in writing speed, at the cost of placing the burden of comprehensibility of the written text on the shoulders of the reader. In any case, the abjad has regularly been judged deficient in rendering the phonology of Arabic.

2.1. Resorting to the Latin Script

A drastic solution is the recourse to the Latin script used phonographically for consonants and vowels.

2. A list of novels written in Egyptian Arabic is to be found in Davies (2005).

3. In 2016, the reference work of Arabic literature *Risālat al-ḡufrān* ‘The Epistle of Forgiveness’ by Abū al-‘Alā’ al-Ma‘arrī (973-1057) was translated from Classical Arabic into Egyptian Arabic by Nārīmān al-Šāmīlī (al-Kutub Ḥān lil-naṣr wa al-tawzī‘, Cairo). In her introduction, she explains that as a young reader, she was unable to understand this classic text, which then motivated her to translate it into Egyptian Arabic, in order to make it accessible to Egyptian readers (Rosenbaum, 2017, pp. 59–61).

TABLE 1. Abjad writing

Non-vocalized abjad writing	#chars
بدون الديمقراطية لا حرية ولا حقوق ولا مساواة. ⁴	45
bdūn -ldīmqrāṭya lā ḥrya wlā ḥqūq wlā msāwā-. ⁵	45
Vocalized abjad writing	#chars
بدون الديمقراطية لا حرية ولا حقوق ولا مساواة.	74
bidūni l-dīmuqrāṭiyyati lā ḥurriyyata walā ḥuqūqa walā musāwā.	62
Without democracy, there is no freedom, no rights, no equality.	

Some contributors to Wikipedia maṣri suggested using what they call “al-frānkō ‘arab,” similar to what is known as “arabizi,” with the use of Latin letters for the corresponding sounds in Arabic, and numbers, or even digraphs, for the others. This is illustrated in Table 2. They highlight the ease of writing on a computer or mobile phone keyboard that does not have an Arabic alternative.

TABLE 2. Wikipedia maṣri examples of experimental articles⁶

“frānkō ‘arab” / “arabizi”
Botros Botros 8aly howa diplomasy Ebti Masry [...]. Kan sekerter 3aam el Omam el Motta7eda el sades [...], we howa 7afed Botros Nayruz 8aly, ra2ees wozara2 Masr fe awa2el el karn el 3eshreen. [...] 5ad el doktorah fel Kanoun el Dawly men Gamey3et Paris [...].
Boutros Boutros-Ghali is an Egyptian Coptic diplomat [...]. He was the sixth Secretary-General of the United Nations [...], and is the grandson of Boutros Nayrouz Ghali, Prime Minister of Egypt in the early twentieth century. He got a doctorate in international law from the University of Paris [...].
“phonetic writing”
El Malti heyya-l loġa-l wataneyya betāeet Malta, we heyya kamān wahda m’n el loġāt er rasmeyya fe-ll Ettehād el Oropi. El loġa-l Malti mettaxda m’n el Ǝarabi w’ be-z zāt el lahgāt betāeet Camāl Afrikya [...].
Maltese is the national language of Malta and one of the official languages of the European Union. The Maltese language is derived from Arabic, particularly North African dialects [...].

4. This is a quote from Syrian poet and essayist Adonis, pronounced at a symposium entitled “Towards a New Religious Discourse” held on February 4, 2015, at the Cairo International Book Fair.

5. The Arabic transliteration system used here is based on the one recommended by the journal *Arabica*.

Others tested in experimental articles a “phonetic” writing with Latin letters, some of them enriched with diacritical marks. The goal is to write Egyptian Arabic by representing the pronunciation more precisely than with the abjad mode of Arabic.

This method is even presented as a “liberation of the Egyptian language from Arabic letters incapable of representing Egyptian speech”⁷.

2.2. Transcribing the Pronunciation of Egyptian Arabic

This attempt to transcribe Egyptian Arabic using graphs from the Latin alphabet is nothing new. In 1897, Daniel Willard Fiske (1831-1904), an American academic, published *An Egyptian alphabet for the Egyptian People*. In this booklet, he presented a transcription model for writing the Egyptian dialect in the Latin alphabet (see Table 3).

TABLE 3. One of Fiske’s example sentences for his Egyptian alphabet

Fiske	wâgib ‘ala kull wâhid jeħibb maşr, in jisâ‘id ‘ala migy el jôm elly fyh kull en nâs es sâkinyn ‘alal nyl jimkinhum jiqru we jiktibu el lisân elly jtkallimuh we jifhamuh. (Fiske, 1897, p. 4)
IPA	wɑːgib ʔala kullə wɑːhid jɪħibbə mɑsˤrə ʔin jisɑːʔid ʔala migi l joːm illi fiː kullə nnaːs issakniːn ʔala nniːl jimkinhum jiʔru wə jiktibu lliːsɑːn illi jtkallimuː wə jifhamuː
English	It is the duty of everyone who loves Egypt to help bring about the day when all the people living on the Nile can read and write the language they speak and understand.

The goal was nothing less than to have this alphabet adopted in Egypt for the teaching of written Egyptian Arabic, which would replace Classical Arabic in everyday use (Zack, 2014, p. 4). Despite the distribution of a whole range of promotional material to Egyptian teachers, this alphabet was never adopted.

Many among such Western scholars linked the high prevalence of illiteracy in Egypt at the time to the diglossic situation of the Arabic language. Since the Arabic abjad was intrinsically associated with Classical Arabic, which was considered complex to master, abandoning it in favor

6. Wikipedia, ويكيبيديا:طريقة الكتابة/ارشيف 1 – [Wikipedia talk: Writing method/Archive 1 – arz.wikipedia.org].

7. Wikipedia, ويكيبيديا:طريقة الكتابة/ارشيف 1 – [Wikipedia talk: Writing method/Archive 1 – arz.wikipedia.org]

of a phonographic system, where vowels would be systematically written, seemed an effective way to allow a greater number of ordinary people to access reading and writing. The association of the Latin alphabet with modernity, and of the Arabic abjad with the past, was a common view at the time (Al-Toma, 1961).

Many other transcription systems have been created since the 19th century to give the most accurate phonographic representation of the pronunciation of Egyptian Arabic, mostly for the needs of foreign learners of the language.

For some of the phonemes, a one-to-one correspondence between Egyptian Arabic phonemes and Latin letters is straightforward and indeed used by all the transcription systems consulted (see Table 4).

TABLE 4. A selection of transcription systems with single graphs

Arabic letters	IPA	Spitta 1880	Gairdner 1916	Mitchell 1956	Louis 2008	Abdel-Massih 2009	Al-Tonsi 2010	Boutros 1993
ب					b			
ت					t			
ج					g			
د					d			
ر					r			
ز					z			
س					s			
ف					f			
ك					k			
ل					l			
م					m			
ن					n			
ه					h			
و					w			

But for other phonemes, a basic Latin graph may be used with a diacritic. This is particularly the case for the emphatic consonants that have their own graph in Arabic and a few other phonemes with no equivalent in the target language of the transcribers. We even find instances of the use of the Arabic graph ‘ayn (see Table 5).

Finally, for a certain number of phonemes, transcriptions vary between basic Latin graphs, sometimes with a diacritic on the one hand or digraphs (and even a trigraph) on the other hand. The first solution allows for a nice isomorphic correspondence to be maintained, but, in addition to the lack of consensus on the choice of diacritics, it forces modern scribes to resort to rare graphs that are beyond the range of the common keyboard. The second solution is the digraph. It is easy to im-

TABLE 5. A selection of transcription systems with symbols and diacritics

Arabic letters	IPA	Spitta 1880	Gairdner 1916	Mitchell 1956	Louis 2008	Abdel-Massih 2009	Al-Tonsi 2010	Boutros 1993
ص	s ^ʕ	ʃ	s	ʃ	ʃ	ʃ	ʃ	ʃ
ض	d ^ʕ	ḍ	d	ḍ	ḍ	ḍ	ḍ	ḍ
ط	t ^ʕ	ṭ	t	ṭ	ṭ	ṭ	ṭ	ṭ
ظ	z ^ʕ	ẓ	z	ẓ	ẓ	ẓ	ẓ	ẓ
ح	h	ḥ	h	ḥ	ḥ	ḥ	ḥ	ḥ
ع	ʕ	ʕ	ʕ	ʕ	ʕ	ʕ	ʕ	ʕ
ء	ʔ	ʔ	ʔ	ʔ	ā ē ō ū	ʔ	ʔ	ʔ

plement and already common in European languages (ch, th, gh, ...) (see Table 6).

TABLE 6. A selection of transcription systems with symbols and digraphs

Arabic letters	IPA	Spitta 1880	Haggenger 1892	Willmore 1901	Gairdner 1916	Mitchell 1956	Louis 2008	Girod 2016
خ	x	ḫ	ch	kh	x	x	kh	ḫ
ش	ʃ	ś	sch	sh	ʃ	ʃ	sh	ś
غ	ɣ	ġ	ġ	gh	ɣ	ɣ	gh	ġ
ا	aː	â	a	â	aː	âa	a	â
و	uː	û	u	û	uː	ûu	uu	û
و	oː	ô	o	ô	oː	ôo	oo	ô
ي	iː	y	i	î	iː	îi	ii	î
ي	eː	ê	e	ê	eː	ée	ee	ê

But the digraph is ambiguous, given that a sequence of two letters can correspond to one phoneme or two phonemes.

In the latter case, the solution adopted by Willmore (1901), is to place a hyphen between the two letters, which looks like an unwieldy quick fix, establishing a separation right in the middle of a word (see Table 7).

TABLE 7. disambiguizing digraphs (Willmore, 1901, p. 1)

Digraph	One phoneme	Two phonemes
<gh>	/ɣ/ ṣughaiyar ‘small’	/gh/ ag-har ‘albino’
<kh>	/x/ âkhir ‘last’	/kh/ dik-ha ‘this’
<sh>	/ʃ/ mishi ‘he left’	/sh/ yis-ha ‘he is distracted’

No consensual solution has therefore really emerged to handily transcribe Arabic into the Latin alphabet. Still, even for some Arabs, the recourse to the Latin script has been seen as a desirable solution, notably to write the vowels of Arabic on the same level as the consonants.

3. Writing the Vowels of Arabic

One of the most famous attempts at using Latin graphs for writing Arabic was the project of ‘Abd al-‘Azīz Fahmī (1870-1951). A member of the Academy of the Arabic Language in Cairo, he was no less than the chairman of the committee on writing and grammar when he presented his project in 1944. One originality was to mix Latin graphs, some with a stroke, and Arabic graphs, a few of which were reversed and linked in a single form to the Latin graphs. But his main concern was writing the vowels, whose absence he saw as the main drawback in the Arabic writing system.

Fahmī	<i>al sayfu aṣḍuq inbāsan min l kutubi</i> <i>fay raddi hi l raddu bayna l riddi wa l laḥibi</i>
Arabic	السيف أصدق إنباء من الكتب في حده الحد بين الجد واللعب
English	The sword is more truthful in informing than books, In its edge is the limit between seriousness and play.

FIG. 1. A sentence in Fahmī's proposed alphabet (Fahmī, 1944, p. 156)⁸

Fahmī admitted that the same verse written in Arabic script was much shorter than when written in the one he suggested, but he found it preferable to transfer the burden of explicitation from the reader to the writer. As expected, Fahmī's proposal created a scandal, and it was never implemented, but it is still remembered as a symbol of blatant hostility toward Islam and Pan-Arabism⁹.

8. This is the first verse of an ode by Abū Tammām (803-845) to Abbasid caliph al-Mu‘taṣim for his victory in the 838 battle of Amorium over the Byzantines.

9. For a detailed account of Fahmī's campaign for the use of Latin graphs, see Shraybom-Shivti (1998, pp. 189–193). For a recent example of a hostile reaction to Fahmī's and others' proposals to write Arabic with Latin letters, see Al-Ṣāwī (2017)

3.1. Chasing the Perfect Phonemic Script?

Yearning for a perfect phonographic script is an understandable undertaking, but probably a vain one. As Florian Coulmas cautions, even a system of phonetic transcription like the International Phonetic Alphabet should not be taken to mean that it incorporates “the ultimate solution to the problem of rendering a continuous stream of sound into a discrete array of graphic signs” (Coulmas, 2003, p. 32) .

This is apparent in the differences that exist between Egyptian Arabic and Standard Arabic. One of the three vowels of the latter is the front unrounded open vowel [a] to which some descriptions of Egyptian Arabic add its back twin [ɑ]. They provide the minimal pair *bāba* ‘the second month of the Coptic calendar’ and *bāba* ‘pope, pontiff, patriarch’ (Badawi and Hinds, 1986, pp. 48–49). Other descriptions prefer in this case to speak of a single vowel and of a series of “marginal” pharyngealized consonants noted by a dot under the Latin graph (Watson, 2002, p. 21).

TABLE 8. back [ɑ] or pharyngealized consonants?

	Badawi, Hinds 1986	Watson 2002
2nd month of Coptic calendar	bāba	bāba
pope, pontiff, patriarch	bāba	ḥāḥa
take care of	barr	barr
be faithful (to a promise)	barr	baṛṛ
my dead	mayyiti	mayyiti
my water	mayyiti	mayyiti

The hesitation in choosing which element of the context, consonant or vowel, should be elevated to a phoneme testifies to the difficulty of finding the holy grail of the “perfect” phonographic script. This should remind us that there is more to speech than a mere sequence of consonants and vowels. An abjad, with its focus on the consonants in the syllable, may offer a better ambit for suprasegmental elements, closer to that of a syllabary.

3.2. Why Bother With Vowels?

The obsession with vowels is also strange for a language like Standard Arabic that basically has just three. Quite regularly, the language hesitates between one or the other, while retaining the same meaning. In any case, the writing does not differ.

TABLE 9. Vowel fluctuation in Standard Arabic

Arabic script	Pronunciation		Translation
	Standard Arabic		
يُحَسِّبُ	/jahsabu/	/jahsibu/	he considers
جَنَازَةٌ	/ʒanaːza/	/ʒinaːza/	funeral
جِصٌّ	/ʒasˤsˤ/	/ʒisˤsˤ/	gypsum
بَزٌّ	/buzz/	/bizz/	teat
يُحْرِقُ	/jahruqu/	/jahriqu/	it burns

Similarly, between Standard Arabic and Egyptian Arabic, the differences are often limited to the pronunciation of vowels.

TABLE 10. Vowel fluctuation between Standard Arabic and Egyptian Arabic

Arabic script	Pronunciation		Translation
	Standard Arabic	Egyptian Arabic	
كَبِيرٌ	/kabi:r/	/kibi:r/	big
صَغِيرٌ	/sˤayɪ:r/	/sˤuɣajjar/	small
تَعِبَ	/taʕiba/	/tiʕib/	he was tired
مَشَى	/maʃa:/	/miʃi/	he walked
يَكْتُبُ	/jaktubu/	/jiktib/	he writes

Writing as an abjad thus allows some room to be left for different interpretations of the written word. This makes it possible to accommodate the various pronunciations of vowels in the different regiolects of Arabic and in the standard variety of the language. The principle of the abjad appears indeed to be particularly compatible with the situation of diglossia that exists in Arabic. By highlighting the consonants, it leaves some considerable slack for the reader to interpret the written word. So what could be felt as a deficiency by a few may have turned out to be a useful compromise for many.

4. The Traditional Abjad Chosen for Wikipedia Maṣri

In any case, for Wikipedia maṣri, the suggestions to write in a phonographic alphabet were not retained, and the abjad mode was preferred. The main arguments were that Egyptians are much more accustomed to

writing and reading it, and the suggested change was perceived as being too bold a step¹⁰.

For the contributors to Wikipedia maṣri, rejecting the Latin alphabet and adopting the Arabic graphs implies the recognition that the language of their project is strongly related to Arabic, despite their desire to prove the contrary. Pragmatism is here the main reason for this decision, because most Egyptians have a better command of the Arabic abjad than of the Latin alphabet. Although this choice may seem logical, it is in contradiction with the stated goal of many early contributors to Wikipedia Maṣri, which was to “liberate” their language from its submission to Arabic, felt as an outside influence from the Arabian peninsula.

This stance is particularly illustrated by the evolution of the terms chosen early on to describe the language of Wikipedia maṣri. First, it is al-luḡa al-‘āmmiyya al-maṣriyya—اللغة العامية المصرية ‘the Egyptian colloquial language’, then it becomes kalām maṣri—كلام مصري ‘Egyptian talk’ and the pejorative term “colloquial” disappears. Finally, it becomes al-luḡa al-maṣriyya al-ḥadīsa—اللغة المصرية الحديثة, that is to say ‘the modern Egyptian language’¹¹, thus placing it in an idealized continuum with the ancient (hieroglyphic) Egyptian language. The goal is to minimize its proximity to Arabic while laying claim to more ancient and prestigious (Pharaonic) roots. The preference for the same abjad mode as Arabic is thus a major concession to the recognition of a certain kinship between this “modern Egyptian language” and Arabic.

5. Writing Arabic Differently

However, the early contributors to Wikipedia maṣri have devised ways to differentiate the script they use from the traditional Arabic script.

5.1. Yāʾ and ʾalif Maqṣūra

In Arabic, the graph yāʾ <ي> most often corresponds to the semi-consonant /j/ or the long vowel /iː/. A similar graph but without the two dots <ى>, used exclusively in the final position, corresponds to the long vowel /aː/. Arabic grammarians call it ʾalif maqṣūra. Egyptians are known for merging the two graphs in one, in the final position, most often the undotted ʾalif maqṣūra.

10. Wikipedia, ويكيبيديا:طريقة الكتابة/ارشيف 1 - ويكيبيديا [Wikipedia talk: Writing method/Archive 1—arz.wikipedia.org]

11. اللغة المصرية الحديثة - ويكيبيديا - Modern Egyptian Language—arz.wikipedia.org

TABLE 11. Fusion of yā' and 'alif maqṣūra in Wikipedia maṣri

	يَاء – yā'	ألف مقصورة – 'alif maqṣūra
Standard Arabic	ي علي /ʕali:/ Ali (proper noun)	ي علي /ʕala:/ on top of
Wikipedia maṣri		ي علي

This national specificity was therefore seized upon by the initiators of Wikipedia maṣri to write any final yā' as a 'alif maqṣūra,¹² thus exploiting the treatment of these two graphs in Egypt¹³.

5.2. 'alif Al-waṣl and 'alif Al-qaṭ'

Another subtlety of Arabic is the difference between 'alif al-waṣl <أ> and 'alif al-qaṭ' <أ>, which could be translated respectively as “fusion 'alif” and “separation 'alif”. These are the two main statuses of the 'alif at the beginning of a word. The 'alif al-qaṭ' is a 'alif that indicates a glottal stop /ʔ/ (the separation) generally marked with a hamza <ء>, as a diacritic above or beneath. The 'alif al-waṣl is a token graph that indicates a liaison (the fusion) with the last vowel of the preceding word. It is not pronounced except at the beginning of a sentence, where it is pronounced as a 'alif al-qaṭ' with a glottal stop. In most texts, the 'alif al-waṣl has no diacritic mark <ا>, which differentiates it from the 'alif al-qaṭ' with its hamza <أ>.

Getting one's 'alifs wrong is one of the most common errors in contemporary Arabic (Attia et al., 2015, p. 13). This may be one reason why the general rule set by the initiators of Wikipedia maṣri is to almost systematically write the initial 'alif without a hamza¹⁴.

12. ويكيبيديا: طريقة الكتابة – ويكيبيديا: How to write—arz.wikipedia.org

13. This Egyptian tendency not to dots the final yā' is illustrated in an article by Egyptian author 'Ādil 'Aṭṭia: “The big problem—for me—started when I began writing. Whenever I sent an article or a story that had a final “yā” without its two dots, the non-Egyptian publisher would put them! And if I put the dots, the Egyptian publisher would delete them!” (Aṭṭia, 2010).

14. ويكيبيديا: طريقة الكتابة – ويكيبيديا: How to write—arz.wikipedia.org

TABLE 12. Fusion of ʾalif al-waṣl and ʾalif al-qatʿ in Widipectia maṣri

	ʾalif al-waṣl	ʾalif al-qatʿ
	fusion ʾalif	separation ʾalif
Standard Arabic	آ ابن ibn son	أم أم ʾumm mother
Wikipedia maṣri	ابن ا	أم ا

5.3. Tāʾ Marbūṭa

A third contextual variant of an Arabic letter is the tāʾ marbūṭa (linked tāʾ) <ة>. In Standard Arabic, it is a graph, always placed at the end of the word, which takes the form of the graph hāʾ <ه>, topped by the two dots of the tāʾ <ت>. It is necessarily preceded by the vowel /a/ (short or long) and is pronounced /t/ followed by the vowel of declension, except at the pause (mostly at the end of a sentence), where it is only pronounced /a/.

TABLE 13. Treatment of tāʾ marbūṭa in Widipectia maṣri

	general rule	ʾiḏāfa (genitive construction)
Egyptian Arabic	/il-luɣa l-ʔiŋgilizijja/ the English language	/luɣit isʕsʕiʔi:d/ the language of Upper Egypt
Wikipedia maṣri:	اللغة الإنجليزیه	لغة الصعيد

In Egyptian Arabic, tāʾ marbūṭa is generally pronounced /a/, but it becomes /(i)t/ at the end of the first element of an ʾiḏāfa (genitive construction). To reflect this difference, the initiators of Wikipedia maṣri have established the following rule: “The tāʾ marbūṭa is written according to its pronunciation: with two dots when it is pronounced /t/, and without dots in the other cases”¹⁴. The omission of the two dots on the tāʾ marbūṭa is a phenomenon that has been observed by Refaat El-Farnawany in his thesis (in German) on Egyptian Arabic as a written language, in which he notes the following: “As far as I know, there is no rule for these dots of—ه. When and where they should be written depends on the writer’s writing habits” (Farnawany, 1981, p. 127). It is equally instructive to note that at no point does he mention the two dots under the final yāʾ, which he systematically writes without dots in the examples of his thesis, as if this rule went without saying.

Thus, the contributors to Wikipedia maṣri have managed to give the script used on their project a specific appearance that helps to differentiate it from Wikipedia al-ʿarabiyya.

6. Graph Frequency Comparison Between Wikipedia Maṣri and Wikipedia Al-ʿarabiyya

In the light of these recommendations, it is interesting to observe the frequency of use of graphs on Wikipedia maṣri (also referred to as “arz”) and compare it with that of Wikipedia al-ʿarabiyya (referred to as “ar”)¹⁵. The files containing the latest version of each article were downloaded for each edition from <https://dumps.wikimedia.org> to obtain `arwiki-20240201-pages-articles.xml` (3.5 GB) and `arwiki-20240720-pages-articles.xml` (11 GB) respectively. The 010 Editor software was used to search for each of the graphs. The graphs are ranked from least represented to most represented on arz in relation to ar in Table 14 that lists occurrences that are under-represented in arz relative to ar, and in Table 15 that lists occurrences that are over-represented in arz relative to ar.

Unsurprisingly, among the least frequent graphs on Wikipedia maṣri are the sukūn, which is used to mark the absence of a vowel between two consonants, and tanwīn diacritics, which are used to mark the flexional ending of indefinite nouns and adjectives. The latter system is inoperative in Egyptian Arabic, except for the /an/ <ْ> flexion of common adverbs such as /ʔabadan/ ‘never’ or /taʔriban/ ‘approximately’. The other flexional endings appear in arz mainly in quotations from the Qurʾān. This /an/ <ْ> inflection is the most frequent in absolute numbers on Wikipedia maṣri, but proportionally the least used in comparison with Wikipedia al-ʿarabiyya. This relative rarity probably results from a bot’s intervention to make it disappear¹⁶, which is not without problems of comprehensibility. The three short vowel diacritics <َ>, <ُ>, and <ِ> remain relatively less rare in arz, in particular to disambiguate the pronunciation of a word, as is also customary in fuṣṣḥā.

The graphs <ث> and especially <ذ> are among those less frequently used on arz compared to ar. This is because they transcribe phonemes not used in Egyptian Arabic. However, the relative rarity of <ذ> compared to <ث> is surprising since the recommendations are to keep it if it is pronounced /ذ/. In addition, <ث> and <ذ> can be used to transcribe

15. Wikipedia differentiates its language editions by the codes assigned to them by the ISO 639-3 standard, i.e., “ar” for Arabic and “arz” for Egyptian Arabic.

16. Bots (short for “robots”) generally make automated changes or actions to maintain Wikipedia’s articles and other pages on Wikipedia.

TABLE 14. Graph frequency comparison between arz and ar—Under-representation in arz

graphs	arz			ar			proportion
	Number	Proportion	Rank	Number	Proportion	Rank	ar/arz
ا	17,263	0.0022%	43	4,027,633	0.1433%	35	233.31
آ	611	0.0001%	47	129,168	0.0046%	44	211.40
أ	1,575	0.0002%	46	235,193	0.0084%	43	149.33
إ	15,490	0.0020%	44	2,063,197	0.0734%	40	133.20
أ	4,596	0.0006%	45	451,382	0.0161%	42	98.21
أ	348,229	0.0444%	36	19,730,202	0.7018%	25	56.66
أ	169,397	0.0216%	38	9,507,807	0.3382%	31	56.13
أ	53,702	0.0069%	40	2,616,609	0.0931%	37	48.72
أ	51,656	0.0066%	41	2,382,296	0.0847%	39	46.12
أ	43,806	0.0056%	42	1,159,890	0.0413%	41	26.48
أ	2,275,450	0.2904%	28	54,292,275	1.9312%	16	23.86
أ	152,032	0.0194%	39	2,868,430	0.1020%	36	18.87
أ	5,646,607	0.7206%	23	100,960,306	3.5911%	10	17.88
أ	1,008,905	0.1288%	31	12,334,500	0.4387%	30	12.23
أ	221,304	0.0282%	37	2,488,371	0.0885%	38	11.24
أ	1,573,673	0.2008%	29	94,74,864	0.3370%	32	6.02
أ	5,017,158	0.6403%	25	28,551,540	1.0156%	23	5.69
أ	1,268,363	0.1619%	30	6,591,327	0.2345%	33	5.20
أ	7,213,905	0.9206%	22	34,717,311	1.2349%	22	4.81
أ	3,226,804	0.4118%	27	15,515,668	0.5519%	27	4.81
أ	9,334,034	1.912%	21	44,379,696	1.5786%	19	4.75
أ	972,875	0.1242%	32	4,427,141	0.1575%	34	4.55
أ	63,034,942	8.0443%	3	275,542,547	9.8009%	2	4.37
أ	12,080,564	1.5417%	19	52,283,490	1.8597%	17	4.33
أ	20,309,171	2.5918%	14	87,907,105	3.1268%	11	4.33
أ	3,299,101	0.4210%	26	13,122,675	0.4668%	28	3.98
أ	23,302,609	2.9738%	12	86,999,405	3.0945%	12	3.73
أ	28,206,045	3.5995%	10	102,078,112	3.6309%	9	3.62
Corpus	783,599,634	100%	47	2,811,387,118	100%	47	3.59

the voiceless [θ] and voiced [ð] dental fricative consonants of foreign words, particularly those containing the sequence <th> in English.

The graph <ظ> is slightly less frequent on arz than on ar, This is probably the result of a balance between the tendency to phonetically render the replacement in Egyptian Arabic of the pharyngealized alveolar fricative [zʕ] by the pharyngealized alveolar occlusive [dʕ] for example with <ظهر> /dʕahr/ 'back' and the opposite phonetic trend with <ظابط> /zʕa:betʕ/ 'officer'.

TABLE 15. Graph frequency comparison between arz and ar—Over-representation in arz

graphs	arz			ar			proportion
	Number	Proportion	Rank	Number	Proportion	Rank	ar/arz
ي	71,586,498	9.1356%	2	255,872,222	9.1013%	3	3.57
و	46,553,694	5.9410%	6	161,904,072	5.7589%	5	3.48
ر	45,472,582	5.8030%	7	156,650,474	5.572%	6	3.44
ن	41,468,036	5.2920%	8	141,732,511	5.0414%	8	3.42
ع	23,291,652	2.9724%	13	79,194,474	2.8169%	14	3.4
ا	47,052,142	6.0046%	5	155,964,726	5.5476%	7	3.31
ز	118,929,723	15.1774%	1	391,184,188	13.9143%	1	3.29
س	5,613,067	0.7163%	24	18,326,115	0.6519%	26	3.26
م	27,247,443	3.4772%	11	85,551,920	3.0431%	13	3.14
ط	55,993,480	7.1457%	4	174,238,132	6.1976%	4	3.11
ك	9,366,993	1.1954%	20	25,908,792	0.9216%	24	2.77
ل	20,292,778	2.5897%	15	55,471,216	1.9731%	15	2.73
ص	16,558,167	2.1131%	16	45,069,246	1.6031%	18	2.72
ح	15,821,354	2.0191%	18	36,108,823	1.2844%	21	2.28
هـ	31,729,113	4.0491%	9	38,498,437	1.3694%	20	1.21
ى	16,225,512	2.0706%	17	12,808,495	0.4556%	29	0.79
يس	493,200	0.0629%	34	29,921	0.0011%	45	0.06
ج	598,199	0.0763%	33	21,100	0.0008%	46	0.04
	456,134	0.0582%	35	14,114	0.0005%	47	0.03
Corpus	783,599,634	100%	47	2,811,387,118	100%	47	3.59

The various spellings of hamza <ؤ>, <ئ> and <ء>, as well as those using ʾalif as a carrier: <أ>, <إ>, as well as the madda <آ>, are relatively less used on Wikipedia maṣri compared to Wikipedia al-ʿarabiyya. This corresponds to the virtual absence of pronunciation of the “glottal stop” /ʔ/ originating from a hamza in the middle and end of words in Egyptian Arabic, and therefore to the avoidance of the etymological spelling in favor of a more phonetic spelling. On the other hand, this points to a lack of popularity of the phonetic solution to mark the “glottal stop” /ʔ/, that in Cairo Arabic replaces /q/ from Arabic fuṣḥā, and the strong tendency to retain the etymological spelling <ق>. Incidentally, this grapheme is almost as frequent on arz as on ar.

Concerning the ʾalif, the rule requiring it to be written without hamza (either above, or below) seems to explain the relative rarity of <ا> and <أ> on arz compared to ar and consequently the slightly higher frequency conversely for the “naked” ʾalif <ا> on Wikipedia maṣri compared to Wikipedia al-ʿarabiyya. If we consider the total occurrences of these graphs on each of the two Wikipedia editions, we obtain a ratio of 3.83

occurrences on ar for each occurrence on arz, close to the ratio of 3.59 between the totals of occurrences of all the characters studied.

Among the relatively much more present graphs on Wikipedia maṣri is <◌◌> which, in this case, corresponds not only to the hā' of the fuṣḥā, but also to the tā' marbūṭa <◌◌>, which the recommendations enjoin to be written without its two dots when it is not at the end of a word in a state of annexation, and thus is pronounced /a/.

As for 'alif maqṣūra <◌◌>, it is the only canonical Arabic graph that has numerically more occurrences on Wikipedia maṣri than on Wikipedia al-ʿarabiyya, despite an overall ratio of 3.59 in favor of the latter. This reflects the tendency prevalent in Egypt of writing any final yā' <◌◌> without its two dots, making final <◌◌> an Egyptian grapheme which, in addition to its value in fuṣḥā of 'alif maqṣūra pronounced /a/, takes on that of final yā' pronounced /i/ or /j/ in Egyptian Arabic.

With the deletion of the hamza over and under 'alif and of the two dots of the tā' marbūṭa and of the final yā', the contributors to Wikipedia maṣri have succeeded in giving the script used in their project a specific appearance that helps to differentiate it from that of Wikipedia al-ʿarabiyya. This tendency to simplify the script is not unlike the common style of handwriting, which regularly omits these marks to increase writing speed. This is reminiscent of the popular roots of Egyptian Arabic, which was originally written mainly by hand. It is also visible in caricatures that have long illustrated Egyptian newspapers.

The rule concerning the 'alif al-waṣl and the 'alif al-qatʿ, as well as the one concerning the tā' marbūṭa were already suggested by Faṭḥi Imbābi (born in 1949) in a sort of post-face entitled "Toward a modern Arabic language" written at the end of his 1994 novel *Marāʿi al-qatl* 'Killing grounds'. He even suggested another simplification by removing the two dots above the qāf <◌◌>, a graph that is pronounced /q/ in Standard Arabic, to clarify its pronunciation as a glottal stop in Cairene Arabic by the use of this graph: <◌◌> (Imbābi, 1994, p. 414). He did not, however, go as far as apply this innovation to the Egyptian Arabic dialogues in his novel, and I have found no trace of it elsewhere. The fact that this graph is deprived of a dedicated key on modern keyboards probably does not help its popularization. What remains to be seen is whether getting rid of all these dots and diacritics will not interfere with the readability of the texts.

7. Al-Zāhir Wa Al-Bātin

In a way, deleting elements of the script harks back to the origins of Arabic writing, when dots were not used to discriminate between letters of the same shape. This contrasts with modern versions of the Qurʾan, which not only include dots but are also entirely vocalized.

This contrast may be related to the opposition between the concepts of al-Zāhir and al-Bāṭin mentioned in the Qur'an (57:3) that could be translated as "the Manifest," and "the Unseen". It is an opposition that permeates many aspects of Arab society, found, for example, in the crescent moon taken up as a symbol of the Red Crescent with the lit side of the moon standing out from its shaded part.

The solution of the abjad reflects these two facets: the possibility to write with a minimum of graphs, and let the reader make his own deductions, or with all the diacritics to provide a maximum amount of information. Paradoxically, the designers of Wikipedia maṣri are thus resorting to traditional strategies of Arab culture, while seeking to differentiate themselves from it.

The script of Wikipedia in Egyptian Arabic, therefore, reflects the tug-of-war between the contributors' desire to differentiate the language used from Standard Arabic and the need to offer texts that remain accessible to Egyptian readers. As in many cases where a script is adapted to an essentially oral language, the trade-off is between respect for the etymology or morphology of the word and the need to transcribe its actual pronunciation. Wikipedia maṣri is a major opportunity to write Egyptian Arabic in many fields to which it has rarely been applied before. It is therefore a particularly interesting candidate in the mission to help establish an orthographic standard for Egyptian Arabic.

References

- Abdel-Massih, Ernest T., Zaki N. Abdel Malek, and El-Said M. Badawi (2009). *A Reference Grammar of Egyptian Arabic*. Washington, DC: Georgetown University Press.
- Attia, Mohammed et al. (2015). "Arabic Spelling Error Detection and Correction". In: *Natural Language Engineering* 25.
- Atṭia, 'Adil (2010). "Goose Yā". In: *Al-ḥiṣwār al-mutamaddin* 3145. عادل عطية، ياء الوزّة!، الحوار المتمدّن
- Badawi, S.M. and M. Hinds (1986). *A dictionary of Egyptian Arabic: Arabic-English*. Librairie du Liban.
- Boutros, Wadie (1993). *Ablan wa sablan, méthode d'arabe égyptien du Caire*. Le Caire: Éditions Dar el-Nashr Hatier.
- Coulmas, F. (2003). *Writing Systems: An Introduction to Their Linguistic Analysis*. Cambridge Textbooks in Linguistics. Cambridge University Press.
- Davies, Humphrey T. (2005). "Dialect Literature". In: *Encyclopedia of Arabic Language and Linguistics* 1. Ed. by Lutz Edzard and Rudolf Erik de Jong, pp. 597–604.
- Diem, Werner (1974). *Hochsprache und Dialekt im Arabischen: Untersuchungen zur heutigen arabischen Zweisprachigkeit. Abhandlungen für die Kunde des Morgenlandes*. Wiesbaden: Franz Steiner Verlag.

- Fahmī, ‘Abd al-‘Azīz (1944). *Latin letters for writing Arabic*. عبد العزيز فهمي، الحروف اللاتينية لكافة العربية. Hindawi.
- Farnawany, Refaat (1981). “Ägyptisch-Arabisch als geschriebene Sprache, Probleme der Verschriftung einer Umgangssprache”. Inaugurale Dissertation sur Erlangung des Doktorgrades der Philosophischen Fakultät II. Friedrich Alexander Universität.
- Ferguson, Charles A. (1959). “Diglossia”. In: *Word* 15, pp. 325–340.
- Fiske, Willard (1897). *An Egyptian Alphabet for the Egyptian People*. Landi Press.
- Gairdner, William Henry (1916). *Egyptian Colloquial Arabic. A Conversation Grammar*. Oxford University Press.
- Girod, Alain and Emad Aziz (2016). *Grammaire de l’arabe égyptien d’aujourd’hui*. Paris: Édition Ellipses.
- Haggenmacher, K. (1892). *Grammatik des ägyptisch-arabischen Vulgärdialektes*. Boehme & Anderer.
- Imbābi, Faṭḥi (1994). *Killing Grounds*. فتحي إمبابي، مراعى القتل. Al-nahr lilnašr wa l-tawzī‘.
- Louis, Samia (2008). *Kallimni ‘Arabi Bishweesh*. Cairo: The American University in Cairo Press.
- Mitchell, T.F. (1956). *An Introduction to Egyptian Colloquial Arabic*. Oxford University Press.
- Rosenbaum, Gabriel M. (2017). “Discussing the use of dialect versus standard Arabic in Lenīn al-Ramlī’s Ḥaṣāwī-skits and other texts”. In: *Jerusalem Studies in Arabic and Islam* 44, pp. 37–66.
- Al-Ṣawī, Muḥammad (2017). *Writing Arabic in Latin letters*. محمد الصاوي، كتابة العربية بالحروف اللاتينية. Noor Publishing.
- Shraybom-Shivtiel, Shlomit (1998). “The Question of Romanisation of the Script and the Emergence of Nationalism in the Middle East”. In: *Mediterranean Language Review* 10, pp. 179–196.
- Spitta, Wilhelm (1880). *Grammatik des arabischen Vulgärdialectes von Aegypten*. Hinrichs.
- Al-Toma, Salih J. (1961). “The Arabic Writing System and Proposals for Its Reform”. In: *Middle East Journal* 15.4, pp. 403–415.
- Al-Tonsi, Abbass, Laila Al-Sawi, and Suzanne Massoud (2010). *Kalaam Gamiil, An Intensive Course in Egyptian Colloquial Arabic*. Vol. 1. Le Caire – New York: The American University in Cairo Press.
- Watson, Janet C. E. (2002). *The Phonology and Morphology of Arabic*. Oxford University Press.
- Willmore, John Selden (1901). *The Spoken Arabic of Egypt*. London: Dalcassian Publishing Company.
- Zack, Liesbeth (2014). “Key to Mass Literacy or Professor’s Hobby? Fiske’s Project to Write Egyptian Arabic with the Latin Alphabet”. In: *Al-‘Arabiyya* 47, pp. 1–19.

Reviving Linear Elamite (2300–1880 BCE)

The Oldest Known Phonetic Writing System

Sina Fakour

Abstract. This paper presents the outcome of a two-year research and design project centered on the Linear Elamite script, an ancient phonetic writing system used in Elam (present-day Iran) at the end of the 3rd millennium BCE. Although first discovered in the early 20th century, Linear Elamite remained undeciphered until 2017, when French archaeologist François Desset began its systematic study. In 2020, Desset, along with Kambiz Tabibzadeh, Matthieu Kervran, Gian Pietro Basello, and Gianni Marchesi, announced the successful decipherment of approximately 95% of the script. This significant achievement repositions Linear Elamite as the earliest known phonetic writing system, transforming our understanding of the origins of writing.

In 2021, as part of the “Missing Scripts” program at ANRT (Atelier National de Recherche Typographique), I began a typographic research project aimed at developing a digital typeface for Linear Elamite. The project involved a detailed analysis of the inscriptions, materials, and engraving tools used, in order to create an inclusive typographic solution. The result is Hatamti, a font family consisting of approximately 400 glyphs in four styles. This paper documents the design process, the visual characteristics of the script, and the rationale behind each design decision, with the goal of supporting archaeologists, linguists, historians, philologists, and type designers in preserving and working with this ancient script.

1. Introduction

At the end of the 3rd millennium BCE, Linear Elamite script gave eternal life to the history of the Elamites. Linear Elamite was one of the scripts used in Elam (located in present-day Iran) to transcribe the Elamite language. It was first discovered at the beginning of the 20th century in Susa, Iran (Álvarez-Mon et al., 2018) and up to now, 43 objects containing Linear Elamite inscriptions have been identified. Linear Elamite inscriptions, containing simple and complex geometric signs, are found on different materials (clay, stone, and metal), and they are named by

Sina Fakour  0009-0009-2429-8968

Atelier national de recherche typographique (ANRT) / American University of Beirut (AUB)

E-mail: sina.fakour@aub.edu.lb

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 247–296. <https://doi.org/10.36824/2024-graf-fako>
ISBN: 978-2-487055-10-0, e-ISBN: 978-2-487055-11-7

alphabet letters (A to Z and A' to R'). François Desset, a French archaeologist working in Iran for years, started the decipherment of LE in 2017. In 2020, François Desset, with Kambiz Tabibzadeh, Matthieu Kervran, Gian Pietro Basello, and Gianni Marchesi, announced that they had deciphered about 95% of the Linear Elamite signs. This decipherment is of importance as it can change our current understanding of the development of writing. It is the oldest known phonetic writing system, showing that the phonetization of writing was achieved from the early stages of its invention. The decipherment has also become a reason for a preliminary Unicode proposal to be made.

In October 2021, I started to work on Linear Elamite as a part of the “Missing Scripts” program at ANRT (Atelier National de Recherche Typographique). The project was guided by François Desset. He is an archaeologist specializing in the Middle East and attached to the CNRS Archéorient team and the University of Tehran.

His presence from the beginning of the project allowed me to have access to the most up-to-date archeological resources and to understand the specific needs of an archeologist working with Linear Elamite.

The purpose of this project is to design an inclusive typographic solution based on comprehensive research and careful analysis of the inscriptions that would make the digital transmission and reproduction of Linear Elamite possible. Among others, it would help archeologists, linguists, philologists, historians and type researchers during the process of redacting and publishing their research. It could also help preserve this ancient script and give it a new life through contemporary media.

In this project I investigated the role of the engraving tool and the material on the quality of the signs, and I tried to achieve a methodology of work that can be applied to similar ancient scripts.

The research and design process took about two years under the supervision of professors Thomas Huot-Marchand, Jérôme Knebusch, Charles Mazé, Émilie Rigaud, and Alice Savoie. The outcome is “Hatamti”. A font family that includes about 400 glyphs in four different styles, which will be released under an open-source license. Another, equally important, outcome of this project is the present paper, which aims to analyze the graphical aspects of the Linear Elamite script and show the arguments and observations behind every design decision.

2. A brief history of Elam

Mesopotamian sources allow us to locate the ancient civilisation of Elam, which covered nearly the entire southern half of the Iranian plateau in the late 3rd Millennium BCE. The Elamite cultural influence extended beyond this region, reaching central Asia, present-day Afghanistan, Pakistan, and the southern shores of the Persian Gulf. But then it was re-

duced to only Susa (Susiana) in the Achaemenid period (539–331 BCE) (Vallat, 1998). Elam was the eastern neighbor of Mesopotamia and was often in conflict with it. Geographically, Elam included more than the actual Khuzestan. It was a combination of the lowlands and the immediate highland areas to the north and east (Matthews, 2022). Elam's particularity compared to Mesopotamia was its confederated governmental structure with complex mechanisms (Álvarez-Mon et al., 2018). Different settled and nomadic populations were united within this structure. The Elamite language, which is extinct today, was used by the Elamites to communicate. The vast majority of scholars believe that it is a language isolate. An isolate is a language that has no relatives, that is, it has no demonstrable genetic relationship with any other language. Language isolates are in effect language families with only one member. The best known and most cited language isolates are Basque, ..." (Campbell, 2010).

The name Elam does not come from the local population. It is a name coined by Mesopotamian scribes. It has reached us from Greek *Ay-lam*, itself borrowed from Hebrew *‘Elām*; the Elamites called their country *Ha(l)tamti/Hatamti*, which the Akkadians rendered *Elamtu* and the Sumerians designated with the ideogram NIM "high, elevated." (Vallat, 1998). We can divide Elam history into four main periods:

- The Proto-Elamite (Susa III/Banesh) period (ca. 3400/3200–2800 BCE).
- The Old Elamite period (ca. 2400–1600 BCE).
- The Middle Elamite period (ca. 1500–1100 BCE).
- The Neo-Elamite Period (1100–539 BCE).

The Linear Elamite writing system was used during the Old Elamite period. The latter consisted of three dynasties: Awan (ca. 2400–2100 BCE), Simaškan (ca. 2100–1970 BCE), and Eparthids or Sukkalmahs. (ca. 1970–1500) (Vallat, 1998).

3. The first phonetic script

As far as we know, writing was invented separately in three different regions: the Near East, China, and Mesoamerica. The earliest writing systems were invented in the Near East in about 3500 BCE and in three regions: Egypt, Mesopotamia, and Elam.

Not much is known about the origins of Egyptian hieroglyphs. But the origins of Mesopotamian and Elamite pictograms can be traced further back to previous forms of counting in the region (Schmandt-Besserat, 2014).

The Elamites used different writing systems to transcribe the Elamite language. The earliest one is Proto-Elamite, which is one of the oldest



FIG. 1. Elam map in 1350 BCE—Ian Mladjov's Resources, <https://sites.google.com/a/umich.edu/imladjov/maps>

known writing systems, and it was used from about 3300 BCE to 2900 BCE (Desset, 2016). The first Proto-Elamite tablets were published by Vincent Scheil, a French epigraphist, in 1900. Although more than 1600 Proto-Elamite inscriptions have been found (about 88% of the tablets are found in Susa), the script remains undeciphered. In regard to its high number of signs, it is considered to be a system composed of signs with both phonetic and logogrammatic values (Desset, 2016).

Scholars do not share the same opinion about the relationship between Proto-Cuneiform and Proto-Elamite scripts. Some scholars argue that Proto-Elamite is derived from Proto-Cuneiform. But others, including Desset F., believe that they have both originated from older forms of communication and then separated paths. Therefore, they can be considered sister writing systems that coexisted. The relationship between Proto-Elamite and Linear Elamite is also a topic of debate. Some scholars believe that Linear Elamite evolved from Proto-Elamite, while others argue that there is not enough evidence to support this claim.

Linear Elamite was used during the Old Elamite period (2300–1880 BCE). Eventually, it was completely replaced by a slightly modified version of Mesopotamian cuneiform, which had been used in parallel with Linear Elamite even earlier (Bagh-e-Bidi, 1931).

It is Walther Hinz, German scholar of Persian and Elamite studies (b. 19 November 1906 in Stuttgart, d. 12 April 1992 in Göttingen) (Schmitt, 2003) who coined the name Linear Elamite and Desset F. used the abbreviation “LE” for Linear Elamite and “PE” for Proto Elamite as a convention (Desset, 2022). From now on I will use the same conventions in this text.

In Fig. 2, you can see a simplified diagram of the evolution of writing in the Near East. To put it more simply, in Fig. 3, you can see the two parallel paths that writing has passed through in the Near East. We can see that the emergence of Linear Elamite script is particularly interesting as the phonetization of writing happened in Elam much earlier than anywhere else. The Elamites could achieve an efficient writing system containing alphabetic and syllabic signs in 2300 BCE, while Sumerians and Egyptians were using logograms.

The LE writing system works with 5 vowels, 12 consonants, and 60 syllables. [Fig. 4] (Desset F. et al. 2022). Most of the signs have one or more variants. These variations are most significant between objects found in the eastern archeological sites and those found in the western sites. About 75% of the signs are shared between eastern and western inscriptions. In Fig. 5 you can see all the signs attested in all inscriptions.

4. Linear Elamite Inscriptions

Like PE tablets, Most of the LE inscriptions are discovered in Susa. A city established in 4800 BCE alternated between Mesopotamian and Elamite power. Up to today 43 objects containing LE inscriptions have been discovered. LE inscriptions, containing simple and complex geometric signs, are found on different materials (clay, stone and metal). Some inscriptions read from right to left, while others read from left to right, and they are named by alphabet letters (A to Z and A' to R').

The inscriptions are found in different archaeological sites across Iran. Some of them have unapproved provenance as they were not found in regulated excavations.¹ Here are the numbers of inscriptions found in each archeological site:

- Susa: 20;
- Marvdasht: 1;
- Konar Sandal/Jiroft: 4;

1. Before the 60's, there were clandestine digging or commercial excavations in Iran done by non-specialists. These people who paid for the excavations could possess half of the finds and sell them on the open market. - David Stronach, “Excavations I. In Persia,” *Encyclopaedia Iranica*, IX/1, pp. 88-94, available online at <http://www.iranicaonline.org/articles/excavations-i>

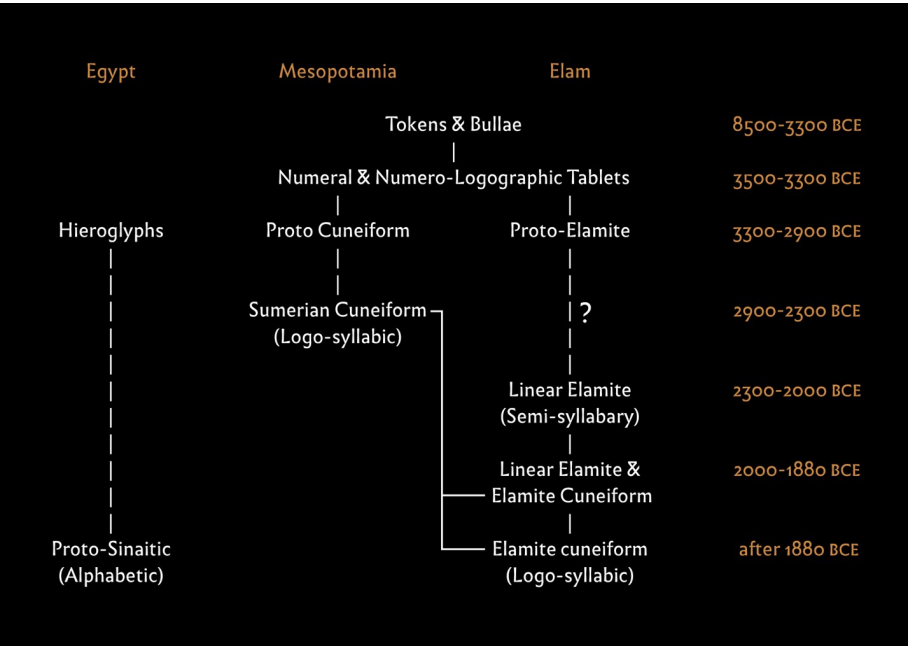


FIG. 2. Diagram of the evolution of writing in the Near East as known today

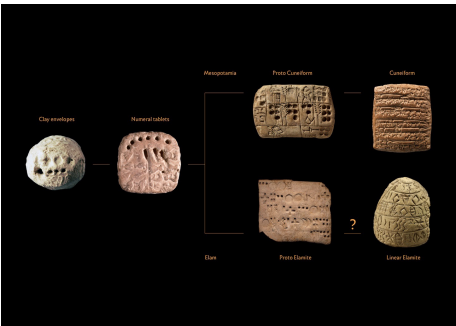


FIG. 3. The evolution of writing in the near east

- Kam firouz: 7;
- Shahdad: 1;
- unapproved: 10.

19 of the inscriptions are kept in the Louvre Museum, 6 are in Iranian museums, 1 is in the Los Angeles County Museum of Art, and the rest of them are in different private collections (Desset, 2022).

	a	e	i	u	u ₂
h ₁ h ₂	ha	?	hi hi ₂	hu hu ₂	
k k ₂	ka	?	ki ki ₂	ku ku ₂	
l	la	?	li	lu	lu ₂
m	ma	me	mi	mu	mu ₂
n	na	ne	ni	nu	
p	pa	pe	pi pi ₂	pu pu ₂	
r	ra	?	ri	ru ru ₂	
s	sa	?	si	su su ₂	
š	ša	še	ši	šu	
t	ta	te	ti	tu	
w	wa	we	?	?	
z	za	ze	?	zu	

FIG. 4. Grid of the most common variants of deciphered signs with their phonetic values. Blue signs are attested in South-Western Iran, red ones in South-Eastern Iran. Black signs are common in both areas. (Photo: Desset F. et al 2022)

The information regarding each inscription (size, material, dating, provenance, writing direction, etc.) is accessible in the article by Desset F. et al 2022.

5. Design process

5.1. Confronting a new script

Translating the engraved signs into contemporary typographic forms and making a consistent system is a challenging process. Designing a digital font for an ancient writing system, which is obsolete today, imposes a significant change in the context of its use. It would therefore be necessary to question our position and define our relationship with the script. If we think of a writing system as the visual representation of a language, how does our visual perception change our understanding of a language? How do we perceive a writing system to which we are foreign? Is our understanding of a new writing system based on the scripts



FIG. 5. Linear Elamite glyph list. (Photo: Desset F. et al 2022)

we already use? Who are the contemporary users of the writing system and what are the contemporary needs? What medium will it be used in?

The first question that arises while working with a different writing system from ours is our legitimacy. In a hypothetical scenario in which I were to work with a living script that had users, I believe it would be crucial to question my position and my relationship to the script. The question remains relevant even if I came from the same geopolitical area, but I was not a member of the community using the script. In that case, getting in touch with the community and informing myself about different aspects of the culture would be necessary.

In the case of LE, the script went extinct long ago, and today there is no community that identifies itself with it. Consequently, the question of cultural appropriation is out of discussion. But being from the same country where the script was used helped me in the research process and in understanding the place the script takes among people and scholars. It also helped me have access to resources in Persian.



FIG. 6. Left: King Puzur-Sushinak inscription, (Table au lion), in two scripts of Akkadian cuneiform and Linear Elamite. Ca. 2120–2100 BCE, Discovered in Susa. Right: Perforated stone, with Linear Elamite text—Département des Antiquités orientales, Louvre Museum.



FIG. 7. Left: Clay cone with Linear Elamite text, dated to the reign of Puzur-Inshushinak. Right: Clay tablet with Linear Elamite text—Département des Antiquités orientales, Louvre Museum.



FIG. 8. Linear Elamite inscriptions on silver objects—Mahboubian collection
<<http://mahboubiancollection.com/>>



FIG. 9. Archeological sites where LE inscriptions were discovered: 1. Susa, 2. Kam Firouz, 3. Marvdasht, 4. Shahdad, 5. Konar Sandal/Jiroft.

Drawing on my past experiences designing Latin and Arabic types, it has become clear that transposing the typographic rules from one system to another would be irrelevant, as they are completely different. In my encounter with LE, I tried to be as neutral as possible and to base my design choices on arguments rather than the typographic rules I am used to. For example, the notion of type color does not seem to be relevant for LE, as we cannot have the same density throughout all signs. Some of the signs consist of numerous strokes that make them dense, and some others have very few strokes.

The main media on which the typeface would be used are paper and screen. This shift in medium, from engraved signs to printed/rendered characters, is considerable. For instance, the engraved signs react to the light in a different way than ink on paper or pixels on screen. So, the stroke thickness should not necessarily be the same (see chapter 5.4).

Another important consideration was to think about the contemporary use of the typeface. The main use of an ancient writing system today would be for academic purposes. The principal users would be archaeologists, linguists, historians, philologists, and type researchers. Having their feedback and considering their needs was at the core of the project. Working with François Desset as one of the main users of this writing



FIG. 10. Excavations to the east of the Achaemenid palace, Susa 1924—Álvarez-Mon, J., Basello, G.P., & Wicks, Y. (Eds.). (2018). “The Elamite World” (1st ed.). Routledge. (photographic archives of the Département des Antiquités orientales, Musée du Louvre).

system allowed me to understand particular needs that the font should meet. (More details in chapter 5.7).

The typeface can also be used by editors, museums, and archives to transcribe the inscriptions. But in some conditions, the sphere of the use of an ancient script can go beyond the academic sphere. Some ancient scripts, such as Cuneiform, have witnessed a wider range of enthusiasts. Some people try to learn how to engrave cuneiform on clay, and others tattoo the signs on their skin. It is not hard to observe the global fascination for a past, considered “grandiose”. There is a wave of nostalgia and desire to revive the glory of the ancestors among nations. The amazing nature of a writing system is that it can become a part of the identity of a community. I cannot neglect this possibility for LE twice as much because of its phonetic nature. Meaning that any language can be written with this writing system. Whether I, as a designer, approve of this usage or not, once the typeface is published, I no longer have control over the usage. What I have control over is trying to assure coherence in the design process and basing my decisions on reason. And eventually provide a typeface with legibility and accuracy. But a certain degree of subjectivity is unavoidable, even in such a rational and methodical approach.

5.2. General analysis of the inscriptions

In the beginning of my research, alongside reading the archeological references, I spent quite a while observing the inscriptions. This played



FIG. 11. Reproduction of Shuwar-asu's clothes on the silver cup. The last king of Iran had a vast program to create an identity based on the ancient history of Iran. A book called *Negar-e-Zan* "Image of woman—An historical pageant of women's dress in Iran" was published in 1973 (Negar-e-Zan, 1973), in which women's clothing on ancient artefacts was reproduced.

an important role in getting familiar with the signs. This was especially helpful in understanding their proportions and skeletons (see chapter 5.5). I also tried to draw them on paper. Despite the good quality of the photos I had from all of the objects (see section 5.7), grasping the real dimensions of the inscriptions could only be achieved by seeing the real objects. A considerable number of the objects containing LE inscriptions are in the Louvre Museum. I managed to closely observe the inscriptions accessible to the public (about 9 objects).

Almost all of the LE signs seem to be abstract and very geometric.² But since they are engraved by hand, they are not completely regular and symmetrical. My first instinct was to design symmetrical geometric signs. An approach that seems to be evident. But was this really the intention of the engravers? Was making perfectly geometric signs desired by the Elamites? Today it seems natural for us to use the word geometric, but were the Elamites aware of this notion?

2. Characterized by or using straight lines, triangles, circles, or similar regular shapes or forms.—Collins, *Geometric* in the Collins dictionary, Retrieved January 15, 2023, from: <https://www.collinsdictionary.com/dictionary/english/geometric>.

To be able to answer this question, I followed a rational approach by trying to answer the five following questions:

- a. Was the notion of geometry known by the Elamites?
- b. Was achieving symmetry desired?
- c. Was achieving consistency desired?
- d. Is there any exceptional inscription?
- e. What are the origins of the signs?

5.2.1. *Was the notion of geometric abstraction known by the Elamites?*

Today it seems natural for us to call a shape geometric after the achievements of the Greeks. Notably because of Euclid's *Elements*³. But was the notion of geometrical abstract shapes comprehended by the Elamites, too?

As preserved in Herodotus History (5th century BCE), the earliest traces of geometrical calculations go back to the Egyptians. They used geometric calculations for their everyday needs, such as religious matters, construction, or timekeeping. Additionally, the Babylonian clay tablets (ca. 1700–1500 BCE) give us the clue that the Pythagorean theorem and some special triads were known way before the Greeks.⁴

But the story of geometric signs goes much further back. Genevieve von Petzinger is a paleoanthropologist and rock art researcher. In her book, *The First Signs: Unlocking the Mysteries of the World's Oldest Symbols*, she studies Paleolithic cave paintings from 40,000 to 10,000 years ago with a particular focus on geometric paintings.⁵

She made a database of more than 5000 geometric paintings from 146 European caves. The result was the observation of 26 signs repeated across different sites over more than 20,000 years. She considers them to be the first traces of graphic communication and symbolism. From the repetition of the same signs in different sites, she concludes the existence of intentionality in drawing these abstract geometric forms. She suggests that the origin of this behavior comes from earlier ages and possibly before our ancestors left Africa.⁶

Although we do not know if these drawings have abstract meanings or not, they do show the capacity of humans to make geometric abstract shapes and reproduce them.

3. Mathematical treatise of 300 BCE, attributed to Euclid. Euclid compiled his *Elements* from a number of works of earlier scientists (Waerden et al., n.d.).

4. Heilbron, J.L. "Geometry". *Encyclopedia Britannica*, <https://www.britannica.com/science/geometry>.

5. Genevieve von Petzinger [WWW Document], 2016. URL <https://www.bradshawfoundation.com/news/archaeology.php?id=Genevieve-von-Petzinger>

6. Genevieve von Petzinger [WWW Document], 2016. URL <https://www.bradshawfoundation.com/news/archaeology.php?id=Genevieve-von-Petzinger>

Ralph Abraham⁷ goes further and argues that the cave abstract geometric forms are proto-geometric primitives and believes that:

“These archetypal patterns that archeologists call geometric forms completed an evolutionary process of abstraction in which an iconic language became a mathematical language” (Abraham, 2015).

Abstract thinking has been a part of human behavior for a long time. Geometric abstraction was one of the first keys in the invention of writing when daily life concepts were translated into geometric markings. Also, these concepts were detached from their context. The emergence of phonetic units shows another level of abstraction where the signs no longer represent real objects (Schmandt-Besserat, 2014).

Therefore, the origins of geometric thinking go way beyond the Elamites. Abstract geometric patterns have been around for tens of thousands of years. The Elamites’ understanding of geometry is probably different from Euclidean calculations, but the existence of these abstract geometric signs on Elamite tablets is not just a simple coincidence but may come from strong intentions.



FIG. 12. Chamber of Spanish tectiforms, El Castillo, Spain—Von Petzinger, Genevieve. “The first signs: unlocking the mysteries of the world’s oldest symbols”. First Atria Books hardcover edition. New York: Atria Books, 2016.

5.2.2. *Was achieving symmetry desired?*

Carving signs on stone or metal is a laborious task, and achieving perfection needs effort. My main question was whether or not the engraver

7. Professor Emeritus of Mathematics at UC Santa Cruz.



FIG. 13. Panel of black Spanish tectiforms, Las Chimeneas, Spain—Von Petzinger, Genevieve. “The first signs: unlocking the mysteries of the world’s oldest symbols”. First Atria Books hardcover edition. New York: Atria Books, 2016.



FIG. 14. The St. Germain-la-Rivière teeth, France. Deer teeth decorated with geometric signs—Von Petzinger, Genevieve. “The first signs: unlocking the mysteries of the world’s oldest symbols”. First Atria Books hardcover edition. New York: Atria Books, 2016.



FIG. 15. The “La Pasiega Inscription,” La Pasiega, Spain. A unique occurrence in Ice Age art—Von Petzinger, Genevieve. “The first signs: unlocking the mysteries of the world’s oldest symbols”. First Atria Books hardcover edition. New York: Atria Books, 2016.



FIG. 16. 26 Non-Figurative Signs repeated in French sites across 25,000 years of prehistory—Von Petzinger, Genevieve. “Geometric Signs & Symbols in Rock Art,” Bradshaw foundation, [WWW Document], n.d. URL https://www.bradshawfoundation.com/geometric_signs/index.php

intended to make completely symmetrical and regular shapes. To answer this question, I followed two different approaches. One based on observation and the other on practical tests.

In the first step, I tried to detach the signs from several inscriptions to be able to analyze them separately. In Figs. 17 and 18, you can see this process for inscription A.

In order to have more material for comparison I chose the signs used the most in all inscriptions which are $\diamond$ (representing the sound a), $\oplus$ (representing the sound r), $\dagger$ (representing the sound t) and the dividing sign $|$ [Fig. 19]. The sign t and the dividing sign do not contain considerable geometric data to analyze. Consequently, I focused on $\diamond$ and $\oplus$. The diamond shape of the sign a is the dominant shape in the LE script. Not only because it represents one of the most repeated sounds, but also because by adding elements to it, some other signs such as $\diamond$, $\diamond$, $\diamond$, $\diamond$, $\diamond$, $\diamond$, $\diamond$, $\diamond$, $\diamond$, $\diamond$, $\diamond$ and $\diamond$ are formed. Also, by removing or adding elements to the sign $\oplus$, we have the signs $\oplus$ and $\oplus$.

I detached these two signs from all of the inscriptions to make the comparison possible. Their sizes vary across the inscriptions. By scaling them to the same size, I tried to remove the proportion factor and just focused on the sign shape. [Figs. 20, 21]

The aim was to verify the existence of any asymmetrical patterns shared between all or some of the signs. For example, if we look at the sign $\diamond$, in inscription C' we can see that the top part of the diamond shape is longer than the bottom part. I then investigated whether this

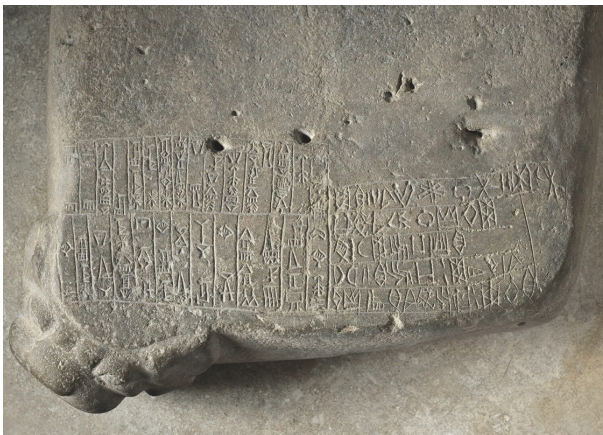


FIG. 17. King Puzur-Sushinak inscription, (Table au lion), in two scripts of Akkadian cuneiform and Linear Elamite. Ca. 2120–2100 BCE, Discovered in Susa—Département des Antiquités orientales, Louvre Museum.

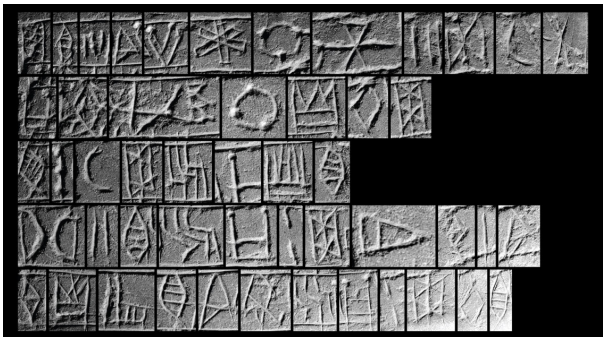


FIG. 18. Detached signs of the inscription A (Table de lion)

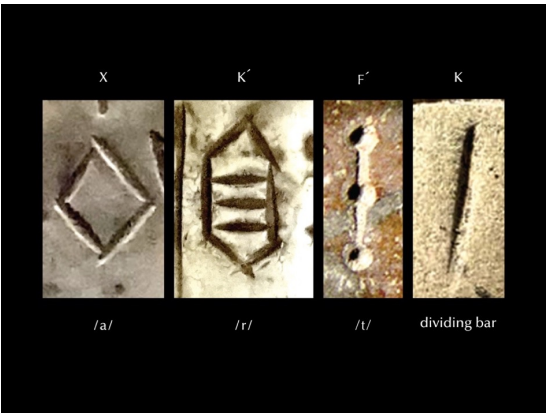


FIG. 19. The signs attested the most in inscriptions



FIG. 20. The sign detached from different inscriptions



FIG. 21. The sign detached from different inscriptions

characteristic was repeated in other inscriptions. The result is that I did not find any asymmetrical pattern shared between more than 2 or 3 signs. The same approach was done with the sign $\hat{\Theta}$ and the result was the same.

The next step was tracing the signs with a line, trying to be as loyal as possible to the inscriptions. [Figs. 22, 23]

I superimposed the drawings to achieve the average shape of the sign $\diamond$. The result (the denser area) shows that the average shape tends to be a symmetrical diamond. [Fig. 24]

Despite the fact that the sign $\hat{\Theta}$ is more complex in terms of geometry, the same result is achieved by superimposing its drawings. [Fig. 25]

Consequently, I could deduce that symmetry is one of the characteristics of the LE signs.

5.2.3. *Was achieving consistency desired?*

The process of designing a typeface includes not only designing individual signs but also making a system in which all the signs work together in

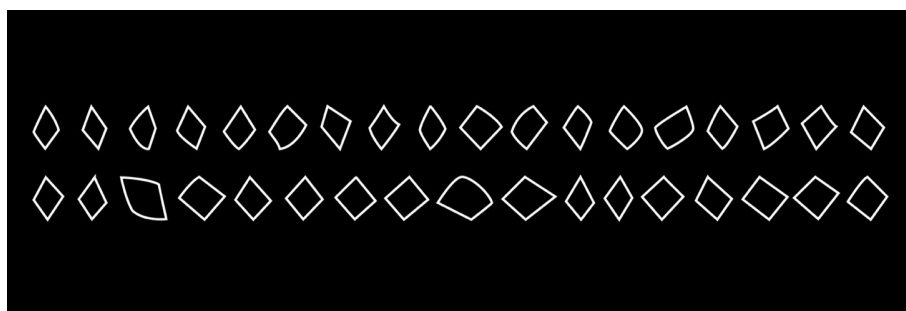


FIG. 22. Tracing the sign ◊ from all inscriptions and scale them into the same height

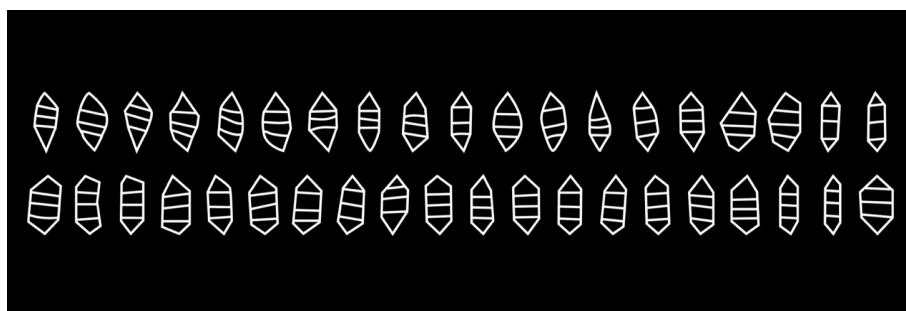


FIG. 23. Tracing the sign 𐎶 from all inscriptions and scale them into the same height

a coherent way. What makes the typeface a system is the characteristics shared between different signs.

But is this idea relevant for one of the first writing systems known so far? Is consistency a requisite for LE to form a writing system?

Defining the word consistency is necessary in order to investigate it. What I describe as “consistency” here is the similarities between different signs that give a homogenous nature to the script. In other words, consistency is “agreement or harmony of parts or features to one another or a whole”.⁸

To investigate this quality, different signs needed to be compared to each other. Meaning I now had to work with lines of text rather than one particular sign.

First, I tried to find signs that appear more than once in each inscription. I observed that the similarities between them are not negligible. As

8. Meriam-Webster dictionary. <https://www.merriam-webster.com/dictionary/consistency>.

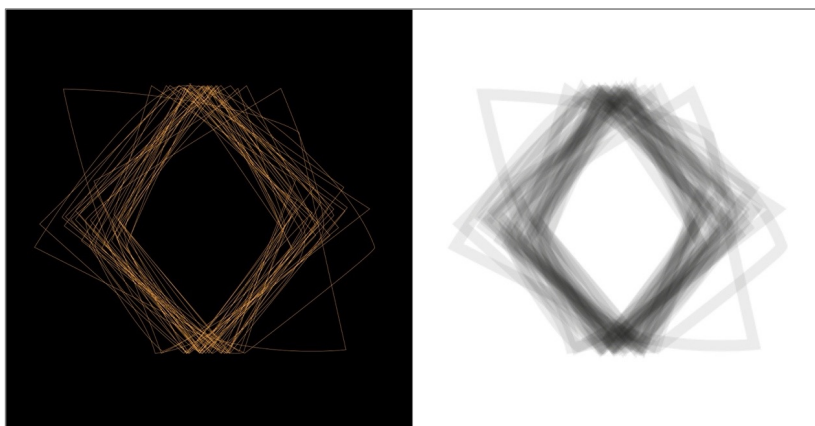


FIG. 24. Superimposition of the skeleton of the signs ◇ from all inscriptions in order to achieve the average shape

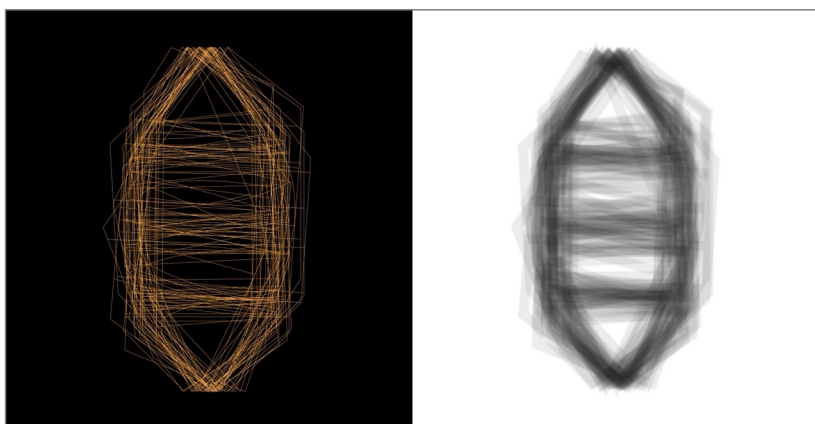


FIG. 25. Superimposition of the skeleton of the signs ⊕ from all inscriptions in order to achieve the average shape

an example, in Fig. 26, object K, you can see the almost identical shape of the sign ◇ attested twice.

But as the number of identical signs in each inscription is limited, to push the argument further, different signs made with similar shapes needed to be investigated too [Fig. 27]. I could also observe a convincing level of shared features between them, making each object coherent as a whole. Keeping the forms similar in repetition is an important element in bringing consistency to the script. A phenomenon that cannot be ac-



FIG. 26. Analysing identical shapes in each inscription

cidental. In Fig. 28, you can see the analysis of both identical and similar forms for various inscriptions.



FIG. 27. Analysing different signs made with similar shapes

5.2.4. *Is there any exceptional inscription?*

As mentioned before, the resemblance between identical or similar signs in each inscription is not negligible. I tend to think that the irregularities can be the result of other factors, such as the engraver's skill. One argument is the existence of the inscription Q or the Marvdasht vessel (21st century BCE). The inscription on this object seems to be meticulously planned. Similar forms are almost identical in terms of shape and size. The spacings are well calculated, the signs are centered vertically between the two lines, and the symmetry is obvious in almost all of the signs.

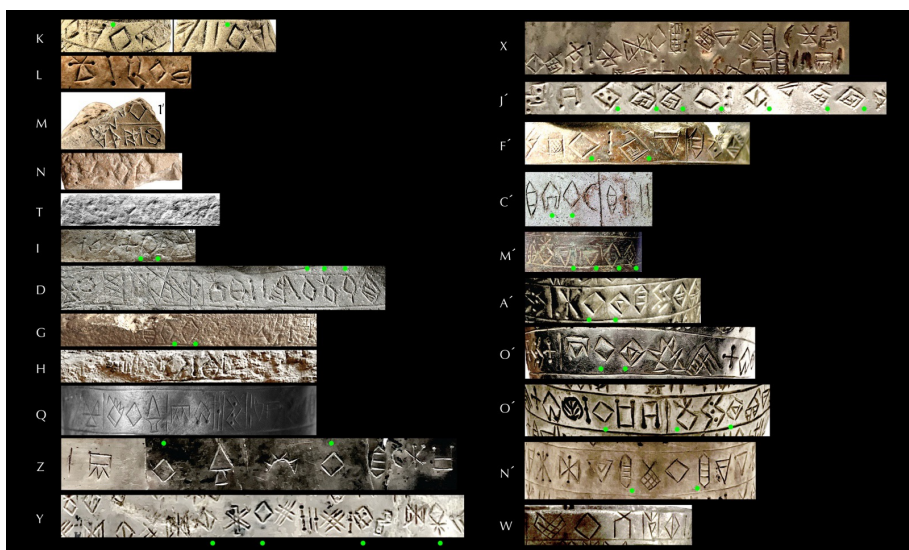


FIG. 28. Analysing identical and similar shapes in each inscription

The existence of the Marvdasht vessel is an important clue about the intention of the engraver to achieve geometric forms that are symmetrical and consistent. [Figs. 29, 30]



FIG. 29. Silver cup from Marvdasht, Fars, with linear-Elamite inscription on the top part. Late 3rd Millennium BCE, National Museum of Iran. (Photo courtesy of François Desset.)

But when we compare similar forms on different objects, despite the same skeleton, we can see more differences in size, position, and relative



FIG. 30. Silver cup from Marvdasht, Fars, with linear-Elamite inscription on the top part. Late 3rd Millennium BCE, National Museum of Iran. (Photo courtesy of François Desset.)

proportions. I believe there are several factors that influence the emergence of these irregularities:

Material

Engraving on deformable materials such as metal or clay is easier than on stone, and erosion affects these materials differently. The inscriptions on metal have remained sharper than those on stone. Also, the engraving space on each object varies, and consequently, the size and proportions of the signs are adapted to each object.

Time and Geography

The area in which LE was used is vast. The inscriptions are found in distant archeological sites and dated within a span of 420 years. This distance in time and place has led to the appearance of different variations for some signs. For example, the sign (representing the sound *n*) is completely different in the inscriptions found in the east and the west of Iran (Desset et al., 2022). [Fig. 31]. Temporal and geographical distance may have also influenced the writing habits. However, this argument needs further investigation with more inscriptions that might be discovered in the future.

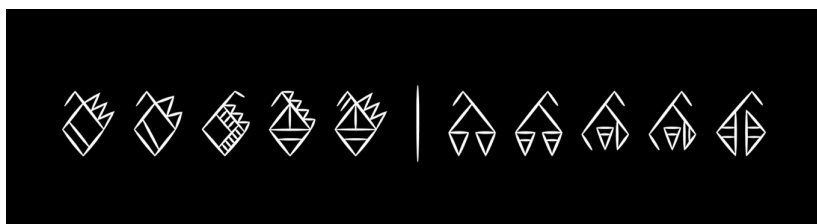


FIG. 31. Variants of the sign representing the sound “n”. Left: Western variants. Right: Eastern variants

Engraver's skill

The engravers had different skills. This can especially be observed by comparing the Marvdasht Vessel with other metal inscriptions of the same period.

5.2.5. *What are the origins of the signs?*

Although there is debate over whether LE is a developed form of the PE script, scholars agree that there is a visual connection between the signs in the two scripts. Most LE signs have potential visual equivalents in PE. What is noticeable in PE inscriptions is the trace of the tool. PE inscriptions are made by pressing a reed stylus on clay tablets. The signs are made by marking their individual strokes one by one, which makes the triangular shape of the stylus wedge noticeable. The round shapes are also done in the same way but with a circular tool.

Most of the PE signs seem to be abstract forms. And even if they have hypothetically represented different kinds of beings or objects, they have been simplified to the level that they no longer have the direct evocative value (Ifrah, 1987, p. 174).

In the process of evolution from PE to LE, the forms become more simplified and stylized. In LE, the presence of the tool is blurred, and we are facing more stylized contours of the shapes rather than embossed imprints. As an example, the round imprint in PE tablets has evolved to just the contour of the circle, and the triangular strokes are simplified to thin lines. Also, the detached strokes become more connected to form integrated shapes.

This transition from 3D to 2D further shows the abstract nature of the LE signs and the fact that they did not mean to reproduce organic objects or beings. Thereupon, it seems relevant not to consider organic forms for the signs and go toward geometric, stylized, abstract forms.



FIG. 32. Proto-Elamite tablet (ca. 3100-2900 BC), Louvre Museum, Paris

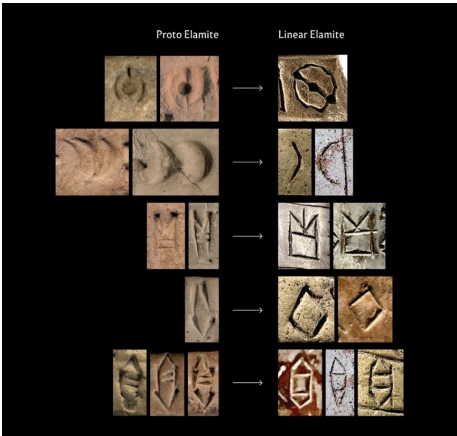


FIG. 33. Evolution of signs from PE to LE

5.3. Analyzing sign details based on the material

When working on an engraved script, the analysis of the materials and tools used can help the designer gain a better understanding of how the forms were shaped. However, there is no information about the engraving tools used to engrave LE, but based on observations, practical tests, and also by looking at early Greek inscriptions, I could make some hypotheses. We should note that LE was developed and used in the Bronze Age when Humans employed stone or clay moulds to shape sophisticated metal tools and weapons.

In the following paragraphs I analyze the quality of the signs based on each material and I investigate what tool could have been used to produce such traits.



FIG. 34. Susa IVA (ca) 2700-2500 BCE), metal objects—Matthews, R., & Fazeli Nashli, H. (2022). “The Archaeology of Iran from the Palaeolithic to the Achaemenid Empire” (1st ed.). Routledge. (image: Javier Álvarez-Mon).

5.3.1. *Metal*

In most of the metal inscriptions, it seems that the strokes were made by putting a blunt chisel-like tool on the surface and hitting it like a punch. In order to have lines with different lengths, one may have used tools with different sizes or used just a part of the width of the tool. Also, in some inscriptions, it appears that the tool was struck a couple of times to extend the line. But what seems more probable is that the lines were not made by scraping or chasing (cutting material by moving forward with numerous small taps). Chasing on metal results in wider endings with accumulation of material and traces of the steps along the stroke [Fig. 36]. For the curved lines, in most cases, we can see the trace of small straight lines joining together to form the curve. This shows that the same technique for straight lines was used, but with a narrower tool. In some instances, the curved lines of the signs are seamless, which suggests the use of a curved tool (Casson, 1935). [Fig. 35]

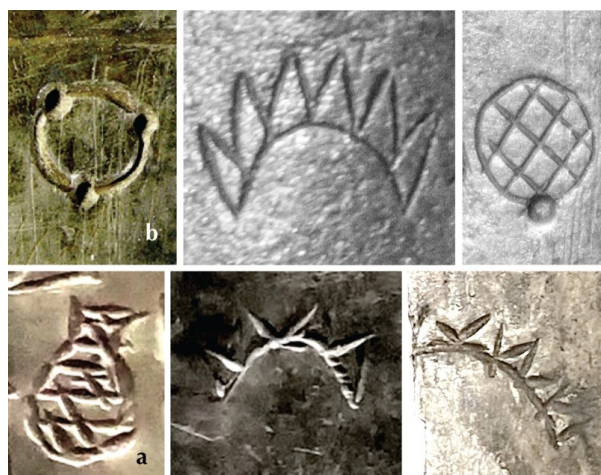


FIG. 35. Quality of the curved lines. In most cases, the trace of several taps is obvious (a), while in a few others the curves are seamless (b)

The strokes are punched separately. Consequently, in their junctions, we see the strokes meet either seamlessly or with a gap.

The other notable phenomenon is that the strokes are the thickest somewhere in the middle, and they fade away at their extremities. [Fig. 36]

This might be due to the fact that the objects have curved surfaces (Casson, 1935). The tool, when put vertically, cannot lie completely flat on the surface. Additionally, the first strokes create a curved surface

around themselves, which accentuates the aforementioned phenomena in secondary strokes. Also, the distribution of the force is not homogeneous, and it is at its minimum in the extremities. The other reason for the thin endings of the strokes could be the use of a round tool that did not have defined edges.

In my experiment with engraving on metal (Zinc), I tried both flat and curved tools [Fig. 37]. A curved tool allows for an easy extension of the line by rotating the tool without removing it from the surface [Fig. 33]. As a result, the long strokes remain seamless, and strokes with different lengths can be done with just one tool. A curved tool also facilitates the rotation of curved signs. With a flat tool, on the other hand, to extend a line I had to remove it from the surface, which caused discontinuity in the stroke. It was also more difficult to achieve curved lines with a flat tool. But I could achieve the best results for the curved lines by using an arc-shaped tool.

The round dots all have the same diameter, showing that they are made by striking a round, pointed tool on the metal.

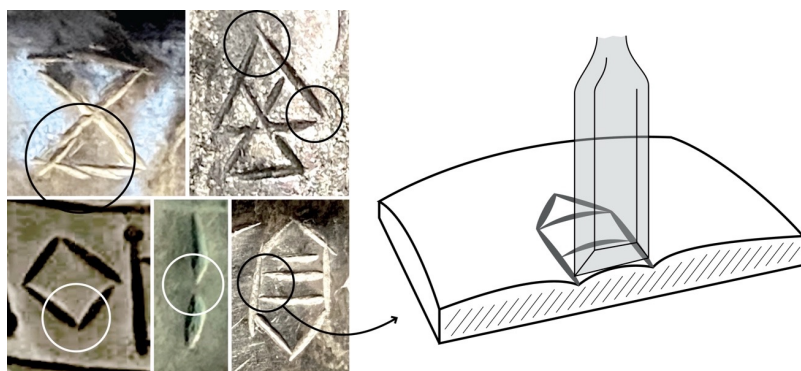


FIG. 36. Left: The quality of straight strokes, their junctions, and their faded extremities in metal inscriptions. Right: Touching points of a flat tool with a curved surface.

5.3.2. *Clay*

Due to the soft nature of clay, signs could be produced by pressing a wide tool (presumably a wooden or reed tool) on the surface, the same way it was done in PE inscriptions. Another method could be scraping using a pointed tool. But this method would cause the accumulation of material at the end of the strokes, which is rarely seen in the inscriptions.

In the case of a tool pressed on clay, what happens is similar to metal, and the stroke extremities look alike. Especially for the secondary



FIG. 37. Experimenting engraving on metal with round and flat tools

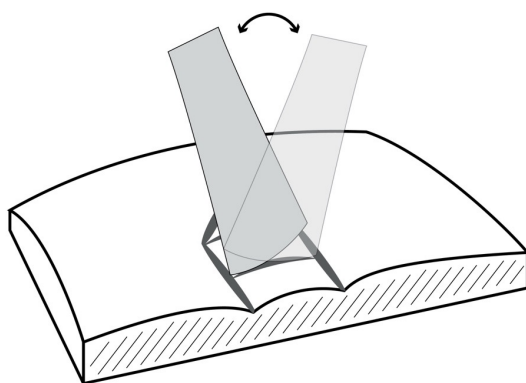


FIG. 38. Extending lines by rotating a curved tool

strokes of a sign, because the first strokes make a cushion-like surface on clay, and the tool in secondary strokes cannot reach everywhere [Fig. 36]. In the case of scraping, the endings would look more defined, which is not the case seen in the inscriptions.

I experimented with clay engraving in practice, and I tried to use three different wooden tools:



FIG. 39. Formation of cushion surfaces and thin extremities in clay inscriptions

Wide rounded tool

The results I achieved with this method are the closest to the inscriptions. In this method, the length of the strokes can be varied based on the degree of pressure and by rolling the tool along its blade. The curved tool makes the strokes thin at the extremities, as the tool does not have defined edges. By pressing the tool deeper, more parts of the curved blade reach the clay, and the stroke becomes longer. Then, by rotating the tool along the stroke to both sides, it can be extended even further. For short strokes, a smaller, similar curved tool can be used. [Fig. 40]

Extension of the strokes can also be achieved by pressing the tool several times along the stroke. Using a wide, rounded tool for engraving is versatile and rather fast.

Wide flat tool

In this method, to alter the length of the strokes, only part of the tool is used. This makes one extremity defined and deep, and the other one, thinner and less profound. Having good control over the length and pressure is more difficult in this method, and the result is less similar to the inscriptions.

Pointed tool

This seems to be the fastest method, but the quality achieved is the farthest to what we observe in the inscriptions. The endings are more defined and have the same thickness as the middle of the strokes. Additionally, material is carried away with the tool, which makes it hard to have defined junctions.

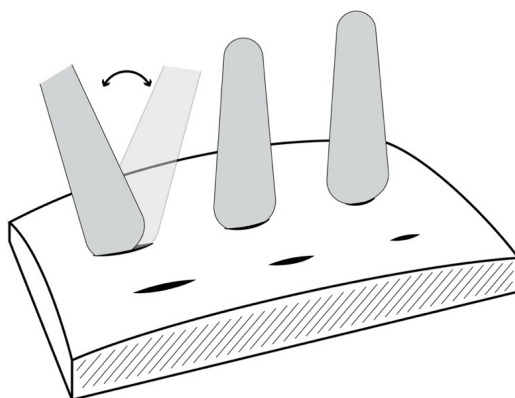


FIG. 40. Thin exterimities in using a rounded tool and how rotation of the tool extends the stroke length

5.3.3. *Stone*

Stone inscriptions (A to I & T to V) are more worn out and less well-preserved, which makes it harder to determine the techniques used. However, the chasing method seems more relevant. In close observation of the objects in the Louvre Museum [Fig. 42] I could see the trace of the taps with which the strokes were carved, probably with a burin-like tool.

However, the thin, less profound lines in softer stones (Such as object C) evoke that the theory of scraping with a sharp tool can also be relevant in some inscriptions. [Fig. 43]

For the curved lines, another possible technique, which was used in Paeonian or early Greek inscriptions, is using a pointed tool vertical to the stone. A laborious work requiring hundreds of punches for each curve (Casson, 1935). [Fig. 44]

Here we can see the same phenomenon of thinner endings in strokes, but less accentuated, and for a different reason. I tried to engrave LE signs on two different stones. One is a soft limestone and the other one has a harder surface. When beginning a stroke, I tended to start with weaker taps so as not to break the stone. In the middle, automatically, I tapped harder to advance more quickly. Reaching the end of the stroke, I reduced the force not to miss the ending. [Figs. 45 & 47]

On the stone, all of the round dots also have the exact same size. From what I have tried, the best way to achieve this is to scrape the stone with a pressing, rotating a pointed tool. Because by hitting the tool (The method used on metal), there would be small breaks on the stone, and we have less control over the perfect shape of the dot.



FIG. 41. Experimenting engraving on clay with different tools



FIG. 42. Trace of the taps in the sign and thin extremities in the sign—Inscription D. Louvre Museum



FIG. 43. Thin lines in inscription C. (Photo courtesy of François Desset)



FIG. 44. Quality of the round shapes. Bottom: Inscription B. Louvre Museum. Top: Inscription on the basis of the Stele of Megakles, ca. 530 BCE, Attica. (Photo: Casson, Stanley. 1935.)



FIG. 45. Top: engraving test on soft limestone. Bottom: engraving test on harder limestone



FIG. 46. Tools used for engraving tests. The left one was used for the dots and the right one for the strokes



FIG. 47. Thin endings of the strokes. Left: Stone inscriptions A & B. Right: Engraving test on soft limestone

After the analysis of different materials, it seemed a relevant decision to keep the concept of the individual strokes, but the question was whether they should be connected or not. Based on what I already discussed in chapter 5.2, I believe that in LE, achieving stylized forms and fading the tool's presence are more important than in PE. This is why I think the desired sign should represent a shape as a whole.

Drawing on what I argued in part 5.2, we can deduce that the inscriptions done more carefully and meticulously are more symmetrical and harmonious. They are those with more homogenous and regular positioning of the signs (like the silver vessel—object Q). We can observe that the more we have these qualities in an inscription, the less we see disconnectivity in the junction of the strokes. This may be another clue to know that the intention has been to make continuous contours. [Fig. 48]

My decision for Hatamti was therefore to keep the contrast between the middle and the end of the strokes, but keep the thin extremities connected. The contrast gives the materiality present in the inscriptions to the typeface. The weight of the connections was chosen in a way so that they wouldn't be lost in small text, and allow the strokes to remain connected in all sizes.

5.4. In pursuit of the suitable weight

Determining the stroke thickness was the other essential question in the early stages of the design process. I found it challenging to define the most suitable value without having at least several signs to compose a text and look at it in print. The main reason for this difficulty is that I was working with a script completely alien to my eyes. No typeface has ever been designed for LE to compare with. So, I could not apprehend how bold or light my drawings are. The only references were the engravings with different line thicknesses on different materials. My solution was



FIG. 48. More detached junctions can be observed in inscriptions with less regularity and harmony. Top: Inscription J', Bottom: Inscription Q (Marvdasht silver vessel)

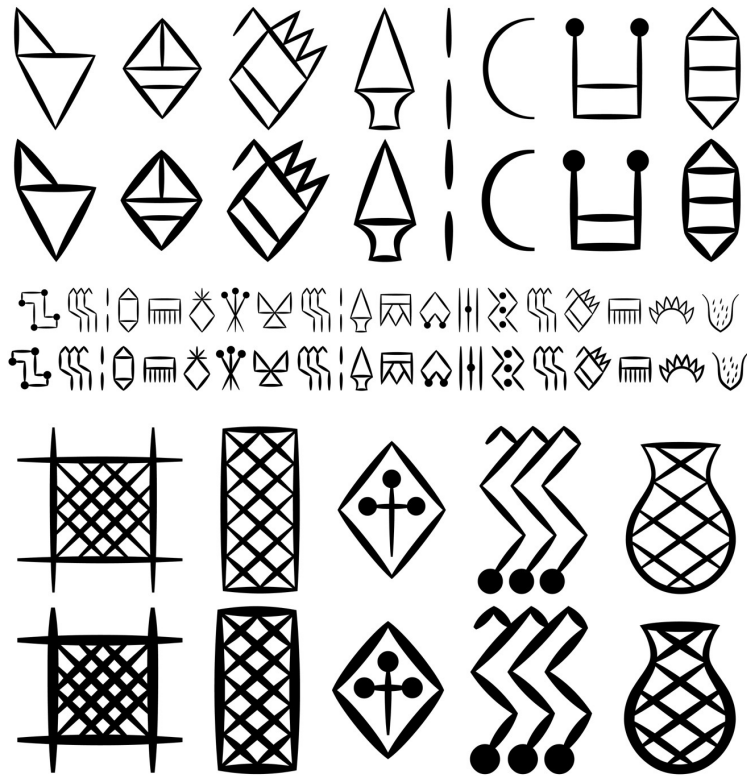


FIG. 49. Hatamti weight in Regular and Bold styles

to draw two different weights as masters and achieve the most relevant weight by interpolating between them later. A necessary but laborious decision, as it doubled the amount of work. This helped me define an intermediate weight as the Regular style for Hatamti. Observation guided this decision: the weight I chose can contain all the complex elements of signs without issues, whilst simultaneously not being so light that we would lose details in print or on screen. It aims to be the closest to the visual perception of the engravings, although it does not have the same thickness. I also decided to export a bolder version for particular uses. The Regular weight I defined for LE is lighter in comparison to text typefaces in modern scripts such as Latin or Arabic (which is the specific nature of this writing system). The Bold version will be useful in pairings with bold texts of Latin or Arabic so that the LE script doesn't appear so weak in the line. It could also be used to emphasize certain signs and establish hierarchy in text.

5.5. Relative proportions of the signs compared to each other

At first glance, it seemed hard to find any pattern for the proportions of the LE signs in relation to each other. The signs share the same shape, but their relative sizes vary from one object to another.

After having the inscriptions under the eyes for quite a long time, eventually some patterns started to merge. While analyzing the most frequent signs ($\diamond$ and $\hat{\text{a}}$), I figured out that in most cases the sign $\hat{\text{a}}$ occupies more vertical space than $\diamond$. This seems logical because if $\diamond$ were as high as $\hat{\text{a}}$, it would be too large and occupy too much horizontal space. [Fig. 50]



FIG. 50. The signs *r* and *a* (or forms similar to *a*), beside each other in different inscriptions

Despite the consistency between the signs of a single inscription, each object has its own accent. In object W, for example, the signs occupy the whole vertical space, which makes them pretty large horizontally. But in object Z, the signs are narrower and have variable heights.

[Fig. 51] We cannot find all of the signs on each object. Some of the signs have even been attested only once and on a single object (like the sign  representing the sound “sa”). Consequently, the main challenge was to establish a harmony between all the signs coming from different objects with different proportions.



FIG. 51. Different sign proportions in each inscription. Wide signs in object W (right) compared to small and narrow signs in object Z (left).

5.6. A system to harmonize the signs

To find a system that harmonizes all of the signs, first I started by drawing the signs  and . As explained in chapter 5.2.b, by adding elements on top, at the bottom, or inside of  we achieve numerous other signs. By reserving some space on top and at the bottom of  I could have the same optical size of this component shared between different signs. By repeating this form, I aimed to add consistency and homogeneity to the typeface. For  and , however, as they should be higher, I drew them in a way to fill these spaces. [Fig. 52]

I applied the same logic to the other characters, and eventually, a system emerged in which we can have four main heights for the signs. [Fig. 53]

In this system, the baseline is positioned in the middle, and the signs are aligned vertically in the center. We can see this clearly in most of the inscriptions and especially in the inscription Q. In about half of the inscriptions, there are lines on top and at the bottom of the signs. We can notice that the smaller signs in these inscriptions are placed in the middle and have the same space above and below [for example, see inscriptions D, Q, A', O', N', and W in Fig. 28].

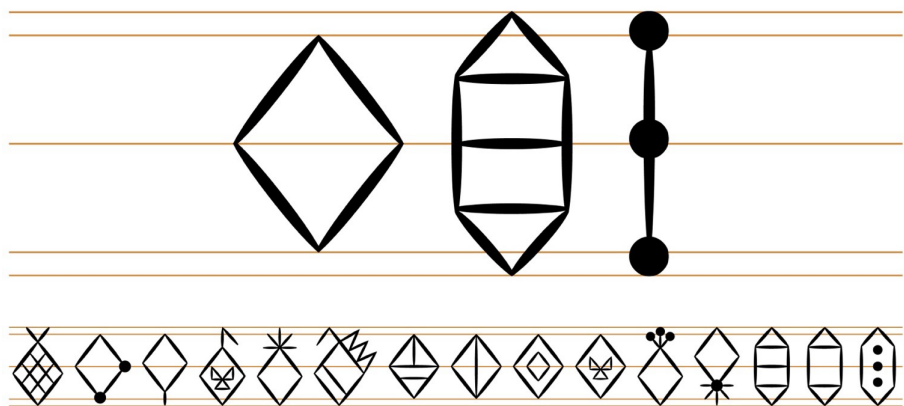


FIG. 52. Defining the proportions of the signs $\diamond$, vertical oval and the signs similar to them

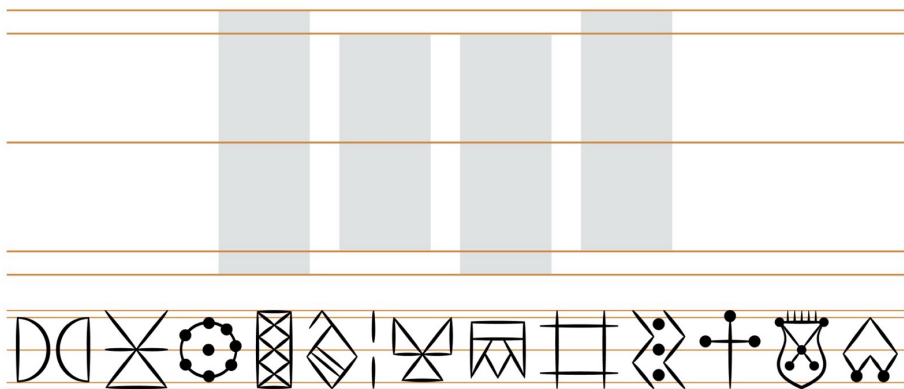


FIG. 53. A system for four main heights of the signs. Baseline positioned in the middle

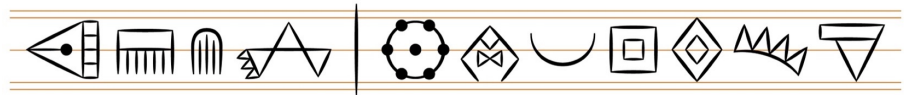


FIG. 54. Some exceptions that do not fit in the grid system

Most signs easily fit into this system. But I did not want the grid to determine the design. Therefore, some smaller signs (about 5% of them) make the inevitable exceptions. The only sign that exceeds the guidelines is a variant of the dividing sign (|), which needed to be distinguishable. [Fig. 54]

The main usage of the typeface being academic papers, I had to consider how it interacts with Latin and Arabic scripts. These two scripts are used the most in academic papers related to LE and the Elam civilisation. For Latin, I did comparisons with Times New Roman as the typeface used the most in academic articles. For Arabic, the comparison was made with Adobe Arabic, Nazanin, and Mitra (the latter two are widely used in Iran for academic purposes).

As LE signs do not have any diacritics or elements added on top or at the bottom, I made the highest signs to occupy all the vertical space from ascenders to descenders. This makes the LE signs taller than Latin characters. The reason is that Hatamti Regular is much lighter than Times, and by making it vertically bigger, it will not appear unnoticeable in the middle of the Latin text. In this way, neither of them dominates the other, but they are easily distinguishable from each other. [Fig. 55]

The same happens with Arabic. The LE signs are bigger to compensate for their lighter weight. Arabic fonts, however, are mostly horizontal and positioned on the baseline. Additionally, the descenders go lower than Latin descenders. This makes LE look positioned higher and somehow on the same baseline as Arabic. [Fig. 56]

5.7. Technical Methodology

In order to begin the design process, I had to observe each sign in detail. Thanks to François Desset, I could have access to high-resolution photos of all of the inscriptions from the early stages of the project. I could also take closer photos of some of the objects during my visit to the Louvre Museum. With the help of the inventory table made by Desset (Desset F. et al 2022), I could trace the objects where each sign can be found.

The process of drawing each sign started with observing it in all its occurrences and comparing them to find the most representative shape. Soon, I understood how difficult it is to control the workflow while having several images open alongside font design software on the screen (if we ignore the music app). So, I printed all of the inscriptions on A3 paper (to preserve the details). For each sign, I put all of the corresponding images beside each other on the table.

I used the Glyphs 3 software to design the font, and I worked with 1000 units per Em scale. To achieve the best proportions, I started with the drawing of the monoline version with a stroke that would form the skeleton of the typeface. Then from the monoline version I drew the

Times

Times New Roman 
 King 

Times Bold

Bold   **Version**
 **Hatamti** 

Times

Line 31 = 25: Whether  or  means “ally, friend” or rather denotes someone hostile is controversial (see EIW 290, s. v. ). The present context seems to require the latter interpretation (but see cun.  and the discussion of the substantive  in Lambert 1972, 64–66). As for the verb, , the context seems to require the meaning “to keep suspended” rather than “to hang (on)” (cf. EIW 1087 f., s.vv. , , and ).

Times Bold

Line 32 = 26:  is generally considered to mean “to scorch” (see EIW 518 f., s.vv. , , and ) but “burned (and) scorched” does not make much sense here. As  appears to pick up , we would expect, by analogy, that  picks up François Desset et al., The Decipherment of Linear Elamite Writing 29  and, therefore, that it has a meaning close to that of the latter.”

Times

Line 31 = 25: Whether  or  means “ally, friend” or rather denotes someone hostile is controversial (see EIW 290, s. v. ). The present context seems to require the latter interpretation (but see cun.  and the discussion of the substantive  in Lambert 1972, 64–66). As for the verb, , the context seems to require the meaning “to keep suspended” rather than “to hang (on)” (cf. EIW 1087 f., s.vv. , , and ).

Times Bold

Line 32 = 26:  is generally considered to mean “to scorch” (see EIW 518 f., s.vv. , , and ) but “burned (and) scorched” does not make much sense here. As  appears to pick up , we would expect, by analogy, that  picks up François Desset et al., The Decipherment of Linear Elamite Writing 29  and, therefore, that it has a meaning close to that of the latter.”

FIG. 55. Hatamti in use with Time New Roman

main style, which is the contrasted one. Therefore, I defined four masters with two axes (weight and optical size). I set the masters' weight values to 500 and 100. The master's optical size values were set to 200 (without contrast) and 100 (with contrast). The weight axis numbers should not be compared to Latin. In Latin, 100 would be considered thin, and 500 would be Medium. However, due to the thin nature of LE, I considered 100 as light and 500 as bold. Then I exported the 250 instance as regular and 500 as bold for the contrasted version. For the monoline version, the instances for regular and bold are 200 and 450. The reason for this difference is that thick junctions and the lack of contrast made the monoline version look bolder, and so I tried to compensate by reducing the values by 50.

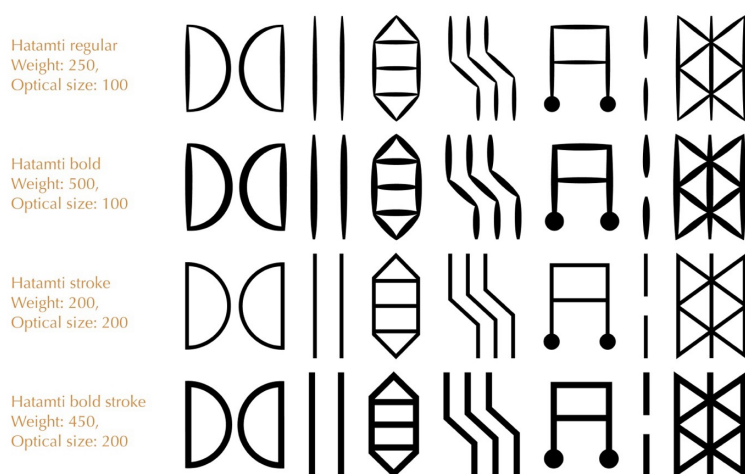


FIG. 57. Weight and optical size values for different Hatamti styles

As LE is not yet in the Unicode Standard, I put the glyphs in the Latin codepoints. LE is an alpha-syllabic system, so I put the signs in the Latin glyphs that have the same phonetic values proposed by Desset F. et al. For the syllabic characters, I used standard ligatures (.liga) of the Latin letters. For example, a “s_i” ligature was made for the sign which has the phonetic value of “si”. In this way, typing in LE with a standard Latin keyboard is possible. Even after the integration of LE in Unicode, keeping this feature would be helpful as it is an intuitive way of typing. Also, the integration of a script in Unicode would not necessarily guarantee the emergence of keyboards for the script. [Fig. 58]

For different variations of each sign, I used the “.cv” suffix so that they could be accessed on different applications. In order to distinguish the sign variants, I made a guide (that would be published alongside the font) in which all the variants of a sign are shown in a table. This document also shows what keyboard keys should be used to access each sign.

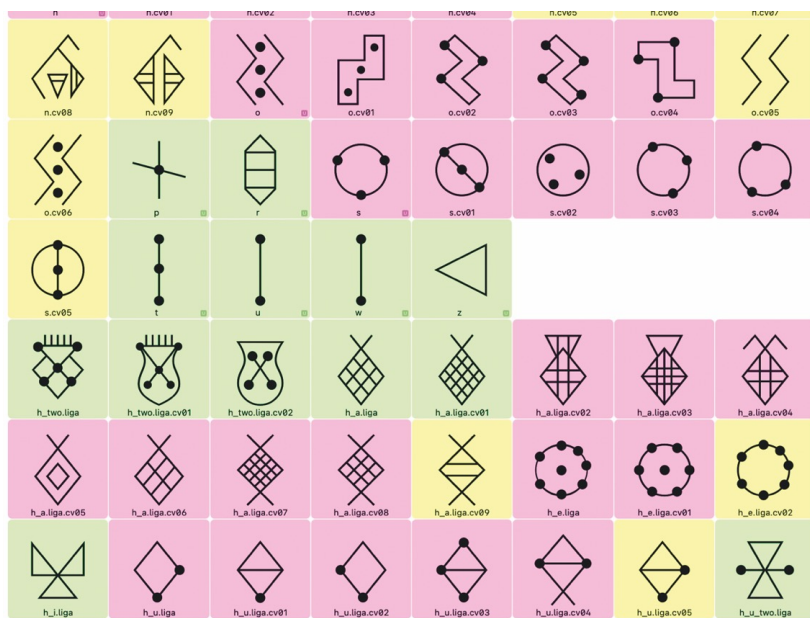


FIG. 58. A preview of the Glyphs application workflow. The glyphs in purple are the western variants, and the glyphs in yellow are the eastern ones. Those in green are common between the east and the west

In transcribing archeological inscriptions, we often face some signs with less legibility due to corrosion and scratches. To be able to represent this, I introduced different hachures representing three degrees of corrosion. By typing an asterisk once after a sign, a hachure superimposes on the glyph. Typing an asterisk twice substitutes the hachure with a denser one. And typing an asterisk three times results in having only the dense hachure without the sign underneath. This feature was achieved by using the “ccmp” feature. same OpenType feature used for Arabic diacritics. This allows for the superimposed hachure on the center of the sign. [Fig. 59]

While reading archeological articles, I observed that some signs like comma, dot, plus, equal, arrow, and dash are used frequently with LE.

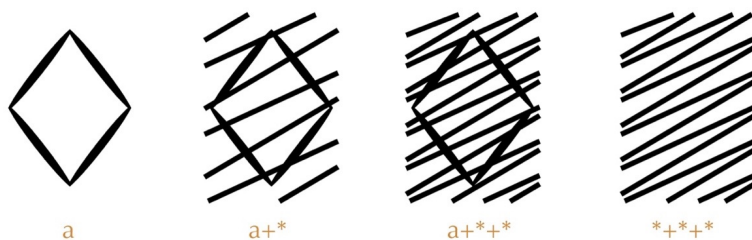


FIG. 59. Scratch superimposition functionality

Therefore, I included them in Hatamti to avoid the need to change the font for such punctuations while typing. You can see the complete glyph-set in Fig. 61.

Before the official release of the typeface, I shared a beta version with François Dasset so that he could try it in practice and report any possible problems. But I also had a couple of meetings with François during the project, in which I presented the drawn glyphs and discussed my design decisions. During our meetings, François provided me with phonetic values of the signs, and we discussed the shape of the signs that are less legible in engravings. Some of the proportions of the signs were perceived differently by each of us, and discussing them better informed my decision-making. As one of the main users of the font, he also expressed his needs in using the font. This led to the implementation of the OpenType features.

5.8. Keeping the skeleton (monoline version)

The monoline (stroke) font was first designed as the skeleton and a way to achieve the main design (contrasted style). But later on, the need to keep it as a separate style was felt by François Dasset. This version represents the skeleton and complete abstraction of the LE signs. The traces of the carving tool and the material are completely erased in this version, and it represents a graphic representation of the signs. Keeping this style gives more choice to the users of the font. Consequently, the final font family has four styles: Hatamti Regular, Hatamti Bold, Hatamti Stroke, and Hatamti Stroke Bold. [Fig. 60]

6. Conclusion

The decipherment of the most ancient phonetic writing system is an important achievement in the history of writing. Making the first LE

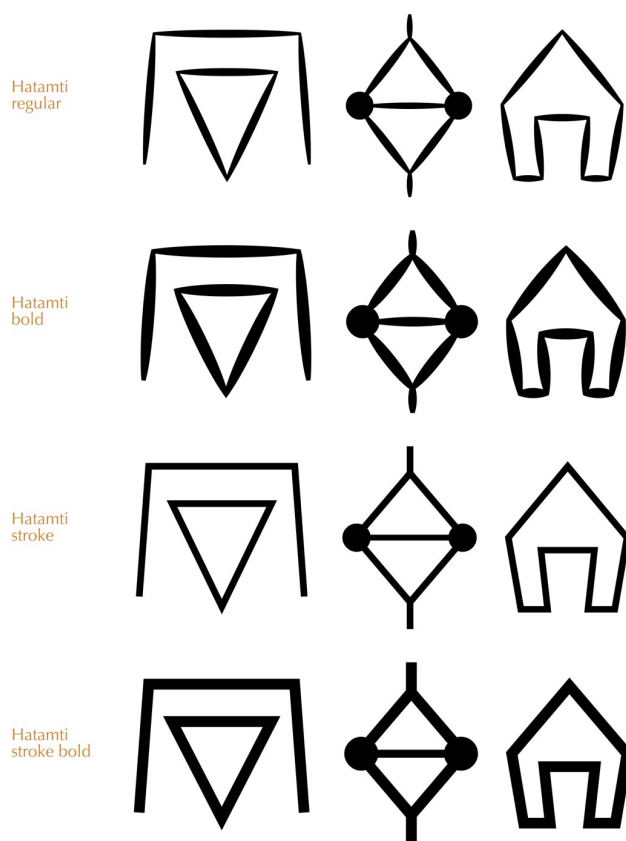


FIG. 60. Hatamti font family

typeface is a delicate task, as it will be the first contemporary visual representation of this writing system. In other words, for the public, LE would be associated with this typeface.

Reviving an ancient writing system requires reconsidering modern typographic principles, including geometry, weight, type color, proportions, contrast, and baseline. These concepts should be evaluated based on the nature of the writing system, but should also be adapted to contemporary media such as print and screen.

In the case of LE, after careful observation and historical analysis, I figured out that achieving consistent, abstract geometric forms is relevant to the nature of the writing system. The regular weight of the typeface is lighter than modern alphabet fonts. Some characters are denser than others, and the type color is not homogeneous everywhere.

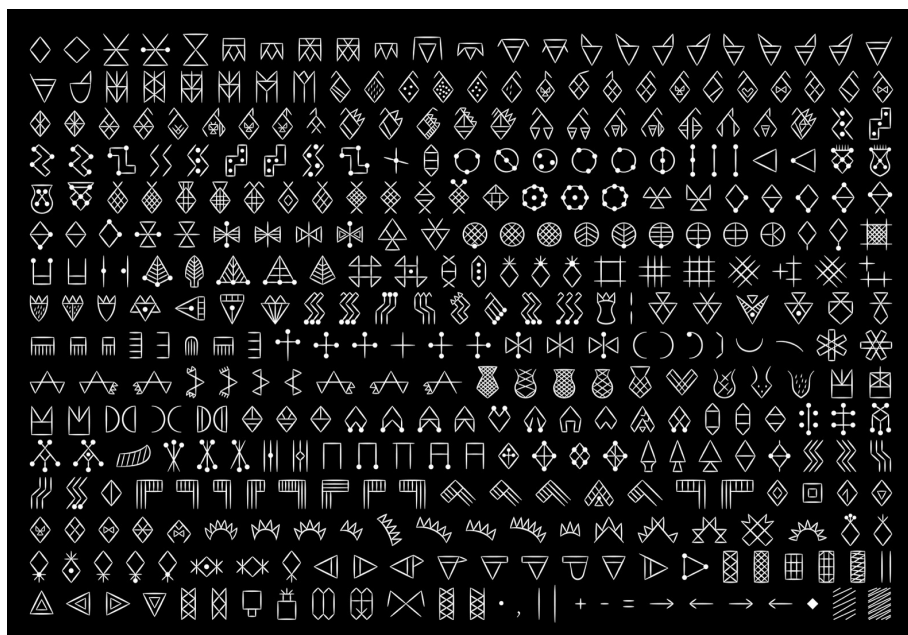


FIG. 61. Hatamti glyph-set

To work on an engraved script, the analysis of the material and tools can help the designer through a better understanding of the forms. This analysis can be done by consulting archaeological references and also by observation, comparison, and tests. In Hatamti, analysis of the tool and material has led to the decision of implementing a contrast and keeping the strokes attached.

Understanding the contemporary users of the typeface and their needs is an important factor in design decisions and the implementation of features. Having François Desset as the archeological advisor made this matter possible.

In the near future, publication of the Hatamti font-family under an open-source license, and its integration in Unicode, would make the digital transmission and reproduction of Linear Elamite possible. Among others, it would help archeologists, linguists, philologists, historians and type researchers during the process of redacting and publishing their research. It could also help preserve this ancient script and give it a new life through contemporary media.



FIG. 62. Four levels of reading an archeological artefact: 1. Object – 2. Inscription photo – 3. Facsimile – 4. Transcription

References

- Abraham, Ralph (2015). "The Paleolithic Birth of Geometric Thinking". DOI: 10.13140/RG.2.1.4819.7846.
- Afshari, Hassan, Yousefi Zeshk, and Ruhollah (2020). "An Analytical Study of the Emergence of Writing in Iran and Mesopotamia in the Late Fourth Millennium BC". In: *Journal of Parseh Archaeological Studies* 11, 27-50. «بررسی تحلیلی پیدایش نگارش در ایران و بینالنهرین در اواخر هزاره چهارم قبل از میلاد»، نشریه مطالعات باستان شناسی پارسه: بهار ۱۳۹۹ - شماره ۱۱ (ISC)، صفحه ۲۷ تا ۵۰.
- Álvarez-Mon, J., G. P. Basello, and Y. Wicks, eds. (2018). *The Elamite World*. Routledge.
- Bagh-e-Bidi, Hassan Rezai (2009). "Calligraphy in Iran". In: *Encyclopedia of the Islamic World*. Vol. 15, 582-592, also in: *Calligraphy and Writing Systems in the Islamic World*, Tehran, 2012, pp. 45-68, «خط، حسن رضائی باغ بیدی، دانشنامه جهان اسلام، جلد ۱۵، ص ۵۸۲-۵۹۲؛ نیز در: خط و نظامهای نوشتاری در جهان اسلام، تهران، ۱۳۹۱، ص ۴۵-۶۸.

- Bennison-Chapman, Lucy E. (2018). “Reconsidering ”Tokens”: The Neolithic Origins of Accounting or Multifunctional, Utilitarian Tools?” In: *Cambridge Archaeological Journal* 29.2, pp. 233–259.
- Bradshaw Foundation (n.d.). *Genevieve Von Petzinger*. URL: <https://www.bradshawfoundation.com/news/archaeology.php?id=Genevieve-von-Petzinger>.
- Campbell, Lyle (2010). “Language Isolates and Their History, or, What’s Weird, Anyway?” In: *Proceedings of the Thirty Sixth Annual Meeting of the Berkeley Linguistics Society*.
- Casson, Stanley (1935). “Early Greek Inscriptions on Metal: Some Notes”. In: *American Journal of Archaeology* 39.4, pp. 510–517.
- Desset, François (2016). “Proto-Elamite Writing in Iran”. In: *Archéo-Nil. Revue de la Société pour l’étude des cultures prépharaoniques de la vallée du Nil* 26.1, pp. 67–104.
- (2018a). “Linear Elamite Writing”. In: *The Elamite World*. Routledge, pp. 397–415.
- (2018b). “Nine Linear Elamite Texts Inscribed on Silver ”Gunagi” Vessels (X, Y, Z, F’, H’, I’, J’, K’ and L’): New Data on Linear Elamite Writing and the History of the Sukkalmah Dynasty”. In: *Iran* 56.2, pp. 105–143.
- (2020a). *A New History of Writing on the Iranian Plateau*. Video. URL: <https://www.academia.edu/video/12472j>.
- (2020b). *Breaking the Code: The Decipherment of Linear Elamite Writing*. Video. URL: <https://www.academia.edu/video/1vyVX1>.
- Desset, François et al. (2022). “The Decipherment of Linear Elamite Writing”. In: *Zeitschrift für Assyriologie und vorderasiatische Archäologie* 112.1, pp. 11–60.
- Heilbron, J. L. (n.d.). *Geometry*. Encyclopedia Britannica. URL: <https://www.britannica.com/science/geometry>.
- Ifrah, G. (1987). *Histoire universelle des chiffres. L’Intelligence des hommes racontée par les nombres et les calculs*. Robert Laffont.
- Matthews, R. and H. Fazeli Nashli (2022). *The Archaeology of Iran from the Palaeolithic to the Achaemenid Empire*. Routledge.
- Pandey, Anshuman (2021). *Preliminary Proposal to Encode Linear Elamite in Unicode*. Script Encoding Initiative. URL: <https://linguistics.berkeley.edu/sei/scripts-not-encoded.html>.
- Schmandt-Besserat, Denise (2014). “The Evolution of Writing”. In: *International Encyclopedia of Social and Behavioral Sciences*. Ed. by James Wright. Elsevier. URL: <https://sites.utexas.edu/dsb/tokens/the-evolution-of-writing/>.
- Schmitt, Rüdiger (n.d.). “Hinz, (A.) Walther”. In: *Encyclopædia Iranica*. Vol. XII. 3, pp. 315–317. URL: <https://iranicaonline.org/articles/hinz-a-walther>.
- Spalinger, Nicolas and Victor Gaultney (2007). *SIL Open Font License (OFL)*. SIL website. URL: <https://scripts.sil.org/OFL>.

- Stronach, David (1999). "Excavations I. In Persia". In: *Encyclopaedia Iranica*. Vol. IX/1, pp. 88–94. URL: <http://www.iranicaonline.org/articles/excavations-i>.
- Swetz, Frank J. (n.d.). *Mathematical Treasure: Mesopotamian Accounting Tokens*. URL: <https://www.maa.org/book/export/html/117930>.
- The International Women's Club of Iran (1973). *Negar-e-Zan: An Historical Pageant of Women's Dress in Iran*. نگار زن : تاریخ مصور لباس زن در ایران.
- Vallat, F. (1998). "Elam I. The History of Elam". In: *Encyclopaedia Iranica*. Vol. VIII/3, pp. 301–313. URL: <http://www.iranicaonline.org/articles/elam-i>.
- Von Petzinger, Genevieve (2016). *The First Signs: Unlocking the Mysteries of the World's Oldest Symbols*. First Atria Books hardcover edition. New York: Atria Books.
- (n.d.). *Geometric Signs & Symbols in Rock Art*. Bradshaw Foundation. URL: https://www.bradshawfoundation.com/geometric_signs/index.php.
- Waerden, Bartel Leendert van der and Christian Marinus Taisbak (n.d.). *Euclid*. Encyclopedia Britannica. URL: <https://www.britannica.com/biography/Euclid-Greek-mathematician>.

Masoretic Graphemics and Graphetics: Challenges and Solutions

Yannis Haralambous

Abstract. Written Hebrew uses two systems of diacritics: vowel marks (*niqqud*) and cantillation marks (*te'amim*). In the Hebrew Bible these marks carry syntactic, accentual, and musical information, and a single accent can decide the meaning of a verse, as the long-disputed reading of Isaiah 40:3 shows. Setting them in type is hard: dozens of sub- and supralinear marks must combine on one line while the base letters keep their normal spacing, each mark stays centered on its letter's vertical axes, and nothing collides with descenders or ascenders. In this paper we first describe the Masoretic writing system as a stack of graphemic layers, define the notion of a graphemic cluster and the graphotactic rules that govern it, and locate the *otiyot meshunot* ("altered letters") in a twilight zone between graphemes and allographs. Tracing one common word across twenty-six printed Bibles from 1517 to 2021, we show that even the finest typography has only partially solved the alignment problem. We then present Tiqwah v2, a corpus-driven system that generates an optimal Masoretic font from a given corpus: diacritic positions are computed by linear-constraint minimization, and the glyph set is kept minimal by a graph-coloring procedure that merges clusters indistinguishable below the perceptual threshold. The method is demonstrated on three fonts of very different character—SBL Hebrew, Monotype Peninim, and the deliberately informal Yarden.

Written Hebrew employs two types of diacritic marks: vowel (or absence of vowel) marks, called *niqqud*, and cantillation marks, called *te'amim*. The former are still used in Modern Hebrew for foreign words or disambiguation, in a way similar to Arabic *tashkeel*, while the latter are specific to the text of the Hebrew Bible (as well as a 19th-century translation of the New Testament into Hebrew) and their use is taught only in religious schools.

Yannis Haralambous  0000-0003-1443-6115

Data Science Department, IMT Atlantique & UMR CNRS 6285 Lab-STICC, Technopôle Brest-Iroise, CS 83818, 29238 Brest Cedex 3, France
E-mail: yannis.haralambous@imt-atlantique.fr

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 297–326. <https://doi.org/10.36824/2024-graf-hara>
ISBN: 978-2-487055-10-0, e-ISBN: 978-2-487055-11-7

1. Introduction

1.1. History of the Masorah

The word *masorah* means “tradition” or “transmission” in Hebrew. The term *Masoretic Text* refers to the corpus of 24 books of the Hebrew Bible as annotated and standardized by the Masoretes, Jewish scribes and scholars active between the 6th and 10th centuries CE. Historically, three competing notation systems have emerged: the Babylonian and Palestinian schools, which used supralinear marks, and the Tiberian system, which was mostly sublinear and combined with a refined vocalization. The Tiberian system eventually prevailed, and the Tiberian Masoretes, especially the Ben Asher and Ben Naphtali families in the 9th and 10th centuries, fixed both vowels and accents on the inherited consonantal text. The Aleppo Codex (c. 930), pointed¹ by Aaron ben Asher, became the touchstone manuscript, and the Leningrad Codex B19a (1008/1009) (see Fig. 1) remains the oldest complete Tiberian witness.

It is widely accepted (Futato, 2020; Park, 2020; Wickes, 1881; 1887; Хаздан, 2020) that cantillation marks serve three primary functions: they segment the text (analogous to punctuation) (Wu and Lowery, 2006), they indicate phonetic stress (Hedánek, 2017), and they prescribe melody (according to various musical traditions) (Kelley, Mynatt, and Crawford, 1998; Levine, 1989). Nevertheless, some scholars, notably Hoop (2008), have contested this view, arguing that cantillation marks constitute predominantly a system of musical notation influenced by other Mediterranean musical notation systems of the period. Mitchell (2017) draws parallels between these marks and the Greek cantillation marks found in a third-century BCE papyrus of a Euripidean tragedy, advancing sixteen arguments for why they could not have been invented by the Masoretes but must rather have been transmitted to them from sources potentially predating the Masoretic period by more than ten centuries.

1.2. Importance of the Masoretic Layer

The following example is paradigmatic of the importance of *te‘amim* as syntactic markers (Futato, 2020, pp. 84–85). In Isa 40:3, there is verse:

1. By “pointed” we mean that the consonantal text was written by a scribe (in this case Shlomo ben Buya’a, around 930 CE), and Ben Asher then added the vowels and cantillation marks on top of that already-written consonantal layer. It was a distinct, highly specialized task—the *naqdan* (pointer) was a master of the oral reading tradition, while the *sofer* (scribe) was responsible for the letters themselves. The division of labor reflects the theological status of the consonantal text, which was fixed and sacred, versus the vocalization, which was understood as a learned tradition laid over it.

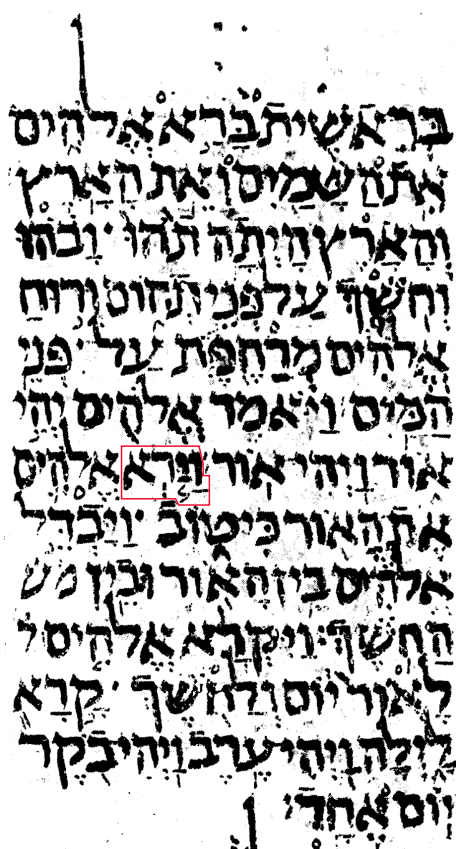


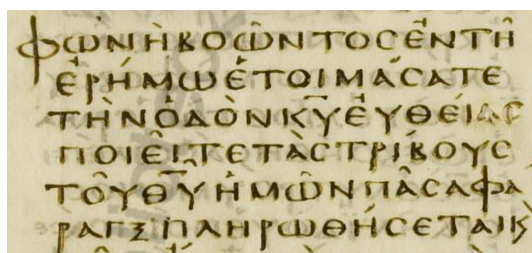
FIG. 1. The first five verses of Genesis from the Leningrad Codex B19a (1008/1009). The word marked is the one used in Fig. 7 to compare 26 printed Bibles from 1517 to 2026.

zaqef gadol → וְ zaqef katan → וְ
 קוֹל קוֹרֵא בַּמִּדְבָּר פְּנֹה דֶרֶךְ יְהוָה

that can be roughly translated as (we purposely do not include punctuation):

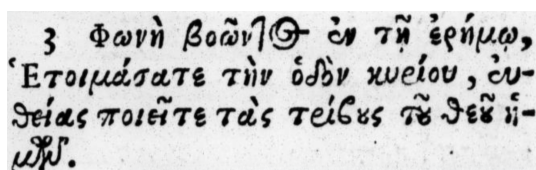
a voice is calling in the wilderness prepare the way for the Lord

This has been translated into Greek and appears in the Septuagint Bible (3rd c. BCE) as shown in this excerpt of the *Codex Vaticanus Vat.gr. 1209* (300-350 CE):



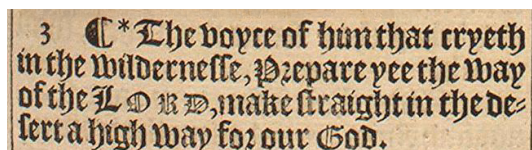
φωνή βοῶντος ἐν τῇ ἐρήμῳ ἐτοιμάσατε τὴν ὁδὸν κυ...

Following 4th-century CE conventions, there are no interword spaces or punctuation. In modern times, punctuation has been added, and we find this passage, in Roger Daniels' *Old Testament* (London, 1653), written as



Φωνὴ βοῶντος ἐν τῇ ἐρήμῳ, Ἐτοιμάσατε τὴν ὁδὸν κυρίου...

where both the comma and the capital epsilon indicate that “in the wilderness” (ἐν τῇ ἐρήμῳ) is syntactically bound to “the voice” (φωνή) rather than to “prepare the way for the Lord” (ἐτοιμάσατε τὴν ὁδὸν κυρίου...). Similarly, in the 1611 King James Bible:



“The voyce of him that cryeth in the wilderness,
Prepare yee the way of the LORD...”

there is a comma after “wilderness,” and “Prepare” is capitalized, as in the Greek text, so that the interpretation is the same as in Roger Daniels.

We could stop the interpretation at this point, if it weren't for the *zaqef katan* cantillation mark on the second word of the Hebrew text, which denotes a primary syntactic division after “calling,” and therefore binds “in the wilderness” not to “the voice” but to “prepare,” leading to the interpretation

A voice rings out: Clear in the desert a road for the LORD,

as it appears in the NJPS Bible edition of the Jewish Publication Society of America (Philadelphia, 1985), which invalidates the “voice calling in the wilderness,” recalled as a prophecy in the New Testament, where

Saint John the Baptist is presented as the personification of this voice (in all four gospels: Matthew 3:3, Mark 1:3, Luke 3:4, and John 1:23). Futato (2020) suggests that this is an error in the Masoretic text, while Hoop (2008, pp. 113–115) is more nuanced.

This example shows that the presence of a specific Masoretic mark suffices to give an entirely different interpretation to a verse that is immensely popular since it is the origin of the idiomatic expression “a voice in the wilderness”.

1.3. The Printed Masora

When Hebrew printing began in late 15th-century Italy, the *niqqud* and *te'amim* posed a serious technical challenge: dozens of marks had to combine cleanly with consonants and vowel points on a single line of metal type. The Soncino family produced the first complete printed Hebrew Bible in 1488, but the decisive achievement was Daniel Bomberg's Rabbinic Bible (*Miqra'ot Gedolot*) in Venice, especially the 1524–25 edition edited by Jacob ben Hayyim, whose pointed and accented text became the printed textus receptus for nearly four centuries. The great Christian polyglots—Complutensian, Antwerp, Paris, London—carried this typography into European scholarship. In the 19th and 20th centuries, critical editions refined the marks: Baer-Delitzsch, Ginsburg, then Kittel's *Biblia Hebraica* (1906; its 3rd edition of 1937 was the first to follow the Leningrad Codex). The *Stuttgartensia* (BHS, 1967–77) and the ongoing *Biblia Hebraica Quinta* (since 2004) reproduce the Leningrad accents diplomatically, while the *Hebrew University Bible* follows the Aleppo Codex.

1.4. The New Testament with Cantillation Marks

Besides the Tanakh, the New Testament has also been translated into Hebrew with Masoretic signs prepared by an English scholar, Ezekiel Margoliouth, and published by the London Society for Promoting Christianity Amongst the Jews, in 1865. The London Society had been producing pointed Hebrew New Testaments since 1821, and a committee revised that text between 1836 and 1838. Margoliouth's was the third revision, and its distinctive contribution was the addition of the full Tiberian apparatus of *niqqud* and *te'amim*, with all the editorial decisions that implies—chiefly, how to map syntactic and rhetorical structures of Koine Greek (parallelism, direct speech, parenthetical clauses) onto the disjunctive/conjunctive accent system designed for biblical Hebrew prose (McClintock and Strong, 1870).

The project's purpose was twofold—missionary outreach to Jewish readers, and a deliberate “naturalizing” of the Greek NT into the lin-

guistic register of the Tanakh. Cantillating the text was the typographic climax of that program: it gave the New Testament the visual and aural signature of Scripture.

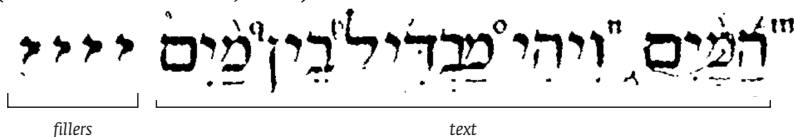
2. Layers of the Masoretic Writing System



FIG. 2. The phrase “[I give] unto him My covenant of peace” (Num 25:12) with colors for the following layers: (1) in black (grey for punctuation), (1') in deep green (broken *vav*), (2) in orange, (3) in violet, (4) in blue, (5) in light green, and (6) in red.

We can define *layers* of elements of a writing system (be they graphemes or non-linguistic) in the following way: two elements belong to different layers if they are not mutually exclusive. Here are the layers we observe (see Fig. 2), inspired by the layers of Arabic (Osborn, 2017):

1. basic letter forms and punctuation;
- 1' special letter forms for the Bible (inverted *nun*, closed *qof* and broken *vav*), and graphe[mt]ic hapaxes (§2.3);
2. a supralinear dot placed on the left or on the right of basic form ש, to distinguish between *sin* and *shin*;
3. the internal dot *dagesh/mappiq* which represents a plosive feature or gemination, or the consonant nature of letter *heh* at word end, and the upper impositive *rafe* that is the binary opposite of *dagesh* and represents the fricative feature;
4. the supralinear short vowel *holam* and sublinear short vowels *niqqud*, half vowels *bataf*, and absence of vowel *sheva*;
5. the sublinear impositive sign *meteg*;
6. *te'amim* cantillation marks;
7. *line fillers* as in the following excerpt of the *Complutensian Polyglot Bible* (Alcalá de Henares, 1517)



2.1. Upper and lower systems

Masoretic text has the unique feature that it can convey not only one but *two* systems of *nikkud* and *te'amim*: the Upper System (*ta'amey ba-elyon*, also called “high cantillation”) and the Lower System (*ta'amey ba-tachton*, also called “low cantillation”). This occurs in texts such as the Decalogue in Exo 20:2–14 and Deu 5:6–18 (Jacobson, 1995), Gen 35:22, and other places. As primary graphemes are the same, and to avoid writing the text twice, secondary graphemes of the two systems are written simultaneously so that a given consonant can have two vowels (the first one always belongs to the high system and the second to the low system) and/or two cantillation marks (see Fig. 3).



FIG. 3. Upper/lower systems

Ashkenazi custom is to use the Lower System for private study and ordinary reading, and the Upper System for public reading of the Decalogue, including the festival of Shavu'ot.

2.2. Graphemic Clusters

Nikkud and *te'amim* can be supralinear/internal/supralinear, prepositive/centered/postpositive (from right to left in the following figure):



These features are graphemic in the sense that a change of positional category can affect the semantics of the text.

More generally, we can subdivide the graphemes used in the Masoretic text into primary and secondary graphemes, the former being stand-alone graphemes and the latter being necessarily “carried” by primary graphemes. A primary grapheme (called “base character” in

Unicode) together with the secondary graphemes it carries is called a *graphemic cluster*. In a cluster, primary graphemes (layers 1 and 1') represent consonants, long vowels, and punctuation, while layers 2-6 are represented by secondary graphemes.

2.3. Graphe[mt]ic Hapaxes

When studying the phonology of a language, we consider that two phones belong to distinct phonemes if we can find two words, differing only by these phones, and having different semantics. We can define graphemes in a similar way and conclude, e.g., that the graphs “i” and “ı” (as in “fi”) represent the same English grapheme ⟨i⟩ since there is no pair of English words differing only by these two graphs and having distinct semantics, while, e.g., they represent distinct graphemes in Turkish. The problem with this statement is that “distinct semantics” is rather vague.

In this section, we describe cases where primary graphemes are transformed in four distinct ways: they can be broken, have their strokes joined, be larger than their surroundings, be smaller than their surroundings, or be suspended. These are called *otiyyot meshunot* (Ginsburg, 1897; Tov, 1999) and belong to several functional categories at once. They are *indexical* (in Peirce’s sense): they do not represent a sound or a morpheme but point at the word, saying “look here, this is special”. They are *critical*: the puncta extraordinaria in particular are now widely understood as the Masoretes’ way of flagging words their tradition regarded as textually doubtful—a kind of pre-modern critical apparatus encoded inside the text rather than alongside it. They are *hermeneutic*: they triggered, and are retroactively explained (in a way similar to the *ḥurūf al-muqaṭṭa’a* in the Qur’an, Bauer, 1921; Nguyen, 2012), by a large body of midrashic interpretation, e.g., the small *alef* of Lev 1:1 read as Moses’ humility; the raised *nun* of Jdg 18:30 preserving the tradition that the original reading was Mosheh but had to be screened to protect the prophet’s name from association with idolatry, etc. And they are *redactional fossils*: in several cases they almost certainly preserve, in petrified graphic form, evidence of earlier textual states—alternative readings, scribal corrections, or the boundaries of redactional units (the inverted *nuns* around the “When the ark went forth” pericope in Num 10:35–36 are the textbook example).

With these explanations in mind, we may raise the question: are *otiyyot meshunot* graphemes or graphetic allographs? It is easy to find minimal pairs (words that differ only by having or not having the *meshunot* feature), but do they carry “distinct semantics”? Depending on the individual interpretations given, they are more or less semantic, so we consider that they belong to a twilight zone between graphemes and allographs. Here is a list, based on Haralambous (1994). In two cases (Est 9:9 and Dan 6:20), we have two *meshunot* in the same word.

2.3.1. *Transformed Letters*

These are: the joined *qof* (Ginsburg, 1897, p. 48) in Exo 32:25 and Num 7:2; the broken *vav* (*loc. cit.*) in Num 25:12. For the sake of completeness we provide an example of the use of the inverted *nun* (*nun hafukba*), which is the only graphe[mt]ic hapax that found its way into Unicode (HEBREW PUNCTUATION NUN HAFUKHA), and which is used as punctuation.

Exo 32:25	אֶהְרֹן לְשִׁמְצָה בְּקִמְיָהֶם
Num 7:2	הֵם הָעֹמְדִים עַל־הַפָּקָדִים
Num 25:12	לֹא אֶת־בְּרִיתִי שָׁלֹם
Psa 107:23	יִיְרְדֵי הַיָּם בְּאַנְיוֹת

2.3.2. *Large Letters*

Gen 1:1	בְּרֵאשִׁית בָּרָא אֱלֹהִים
Gen 5:1	זֶה סֵפֶר תּוֹלְדֹת
Gen 30:42	וּבַהֲעֵטֶרָה הִצָּאן לֹא יִשְׁיִים
Gen 34:31	וַיֹּאמְרוּ הַכְּזֹנֶה יַעֲשֶׂה
Gen 50:23	בְּנֵי שְׁלֹשִׁים גַּם
Exo 34:7	נִצֵּר חֹסֶד לְאֻלָּפִים
Exo 34:14	לֹאֵל אֶתֶר כִּי
Lev 11:42	הוֹלִיף עַל־גֹּחַוֹן וְכָל
Lev 13:33	וְהִתְגַּלַּח וְאֶת־הַנֶּתֶק לֹא
Num 14:17	וְעִתָּה יִגְדֵּל־נָא כָּח
Num 27:5	מַשֶּׁה אֶת־מִשְׁפָּטָן לִפְנֵי
Deu 6:4	שְׁמֵעַ יִשְׂרָאֵל יְהוָה
Deu 6:4	אֱלֹהֵינוּ יְהוָה אֶחָד
Deu 11:21	כִּימִי הַשָּׁמַיִם הָאָרֶץ
Deu 22:6	יִקְרָא קֹן־צִפּוֹר לִפְנֵיךָ

Deu 29:27	גָּדוֹל וַיִּשְׁלַחְכֶּם אֶל-אֶרֶץ
Deu 32:6	הַלִּיהוָה תִּגְמְלוּ-זֹאת
Deu 33:29	אֲשֶׁרִידָה יִשְׂרָאֵל מִי
Rth 3:13	לִינִי הַלַּיְלָה וְהָיָה
Jos 14:11	מִשֶּׁה כָּכָחִי אֲזוּ
2Ki 17:31	עָשׂוּ נִבְחָז וְאֶת-תַּרְתֵּק
1Ch 1:1	אָדָם שֵׁשׁ שָׁנָה
Est 1:6	חֵזֶר כַּרְפָּס וּתְכֵלֶת
Est 9:9	וְאֶת אֶרְדֵּי וַיִּזְתָּא
Est 9:29	וְתִכְתֹּב אֶסְתֵּר הַמַּלְכָּה
Job 9:34	מֵעַלִי שְׁבֻטוֹ וְאִמָּתוֹ
Psa 80:16	וְכִנֹּה אֲשֶׁר-נִטְעָה יְמִינָהּ
Psa 84:4	וּדְרוֹר קָן לָהּ
Pro 1:1	מִשְׁלֵי שְׁלֹמֹה בֶן-דָּוִד
Ecc 7:1	טוֹב שֵׁם מִשָּׂמֶן
Ecc 12:13	סוֹף דְּבַר הַכֹּל
Sng 1:1	שִׁיר הַשִּׁירִים אֲשֶׁר
Isa 56:10	צִפּוֹ עוֹרִים כָּלָם
Dan 6:20	בְּשִׁפְרָפְרָא יָקוּם בְּנִינְהָא
Mal 3:22	זָכְרוּ תּוֹרַת מֹשֶׁה

2.3.3. Small Letters

Gen 2:4	וְהָאָרֶץ בְּהִבְרָאָם בְּיוֹם
Gen 23:2	לְסָפֹד לְשָׂרָה וּלְבִכְתָּהּ
Gen 27:46	אֶל-יִצְחָק קָצַמְתִּי בְחַיִּי

Lev 1:1	וַיִּקְרָא אֶל־מֹשֶׁה וַיְדַבֵּר
Lev 6:2	עַל־מִזְבֵּחַ עַל־הַמִּזְבֵּחַ
Deu 9:24	מִמֶּרִים הֵייתָם עִם־יְהוָה
2Sa 21:19	אֶלְחָנָן בֶּן־יֵעָרִי אֲרָגִים
Est 9:9	וְאֵת אֲרָדִי וַיִּזְחֹק
Job 7:5	רָמָה וְגִישׁ עָפָר
Job 16:14	יִפְרָצֵנִי פֶּרֶץ עַל־פְּנֵי־פָרֶץ
Job 33:9	בְּלִי־פֶשַׁע חָף אֲנֹכִי
Pro 16:28	מִדּוֹן וְנִרְגָּז מִפָּרִיד
Pro 28:17	אָדָם עֹשֶׂק בְּדָם־נַפֶּשׁ
Pro 30:15	בְּנוֹת־הַבִּבְבָּ
Isa 44:14	נֹטַע אֶרֶז וְגִשֵּׁם יִגְדֵּל
Jer 14:2	לְאֶרֶץ וַיִּזְחֹק יְרוּשָׁלַם
Jer 39:13	רַב־טַבָּחִים וְנִבְשׁוּבָו רַב־סָרִיס
Lam 1:12	לֹא אֵלֵיכֶם כָּל־עֲבָרִי
Lam 2:9	טַבְּעוּ בְּאֶרֶץ שְׁעָרֶיהָ
Lam 3:36	לְעֵיֹת אָדָם בְּרִיבּוֹ
Dan 6:20	בִּשְׁרָפָהּ יָקוּם בְּנִגְהָא
Nah 1:3	יִהְיֶה בְּסוּפָהּ וּבְשַׁעְרָהּ

2.3.4. *Suspended Letters*

Jdg 18:30	בֶּן־גִּרְשָׁם בֶּן־מְנַשֶּׁה הוּא
Job 38:13	וַיִּנָּעְרוּ רַשְׁעִים מִמָּנָה
Job 38:15	וַיִּמָּנַע מִרַשְׁעִים אוֹרָם
Psa 80:14	חֲזִיר מִיֵּעַר וְזִז

2.4. Graphotactic Rules

Let us compare two musical notation systems (the Byzantine Greek system and the Armenian *kbaz* system) with the system of Masoretic cantillation marks.

Both the Greek and the Armenian systems place neumes on a separate line above the text.

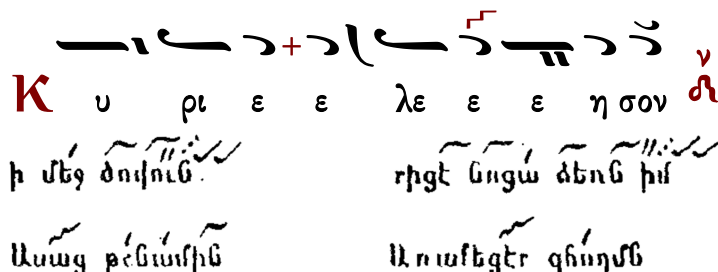


FIG. 4. Samples of Byzantine Greek (above) and Armenian (below) systems of musical notation

In the Byzantine Greek system (Troelsgard, 2011; Ἀθανασόπουλος, 1950) neumes are significantly larger than the corresponding text syllables so that the latter are necessarily spaced to be vertically aligned (and therefore temporally synchronized) with the neumes. In the Armenian *kbaz* neume system (At'ayan and Nersessian, 1999), neumes are sparser and smaller, so that words do not necessarily need to be spaced. Also, the precision of alignment between syllable and neume does not seem to be of paramount importance.

Not so for Masoretic text, where cantillation is represented by secondary graphemes that are “carried” by specific primary graphemes, and therefore their alignments have to be extremely precise.

Subscripts can be horizontally centered or prepositive (right-aligned), while superscripts can be horizontally centered, prepositive, or postpositive. Position is a graphemic feature since, in several cases, visually similar marks denote different signs when changing position (e.g., for example, a centered sublinear “<” denotes mark *mahapakh*; in contrast, in the prepositive position, the same graph denotes mark *yetiv*).

Primary graphemes have vertical accentuation axes for upper and lower diacritics (see Fig. 5). Sometimes, the same consonant can have two lower accentuation axes intended for different marks (see rule 8). For placing the various marks, the following rules have to be followed in the given order:

- Rule 1: supralinear *shin/sin* dots and internal *dagesh/mappiq* marks are placed at invariant positions (at least in printed texts);
- Rule 2: primary graphemes have (sometimes distinct) supralinear and sublinear *vertical axes* (Fig. 5), e.g., in the cases of א, ל, נ, the sublinear axis is their vertical stroke;
- Rule 3: sublinear *niqqud* and “centered” *te’amim* are horizontally centered with respect to these axes: notice the position of diacritics in א, ל, נ, vs. ב, ז, ד, etc.,;
- Rule 4: the *bolam* vowel is a postpositive superscript—an exception occurs for letter *vav* that sometimes carries a prepositive *bolam* mark (*bolam baser*) and sometimes a postpositive one (*bolam male*);
- Rule 5: the subscript *meteg* mark is either centered or follows an existing sublinear vowel—in some cases it may appear between the two components of a semi-vowel;
- Rule 6: in case there is both a sublinear *niqqud* and one or more sublinear *te’amim* of the centered type, the *niqqud* is centered with respect to the vertical axis (see Rule 2) and the *te’amims* are concatenated to its right;
- Rule 7: prepositive and postpositive cantillation marks are placed on their respective positions, and if there is a postpositive *bolam* vowel and a postpositive cantillation mark, the latter will always follow the former;
- Rule 8: when the last syllable of a word has a א, נ or ז consonant with *patah* and is preceded by a long vowel, then the *patah* becomes prepositive and is called *patah furtivum*;
- Rule 9: secondary graphemes must be equally distanced from each other and must not cross or touch descenders (letters *qoph*, *ayin*, and final *kaf*, *nun*, *pe*, and *tsadi*) or ascenders (letter *lamed*);
- Rule 10: most importantly, *the graphs of primary graphemes are set at their regular distances, as if there were no secondary graphemes.*²

It should be noted that the properties of centrality or pre-/postpositivity of *te’amim* may depend on the context. Indeed, some cantillation marks appear twice in the same word: once at its beginning or end, in prepositive or postpositive form, and once on the stressed syllable, in *centered* form—these are called *anticipatory* marks.

2. This was the case in manuscript Bibles since, as already mentioned in note 1, a manuscript Masoretic text was produced in two steps: first the primary graphemes were written by the *sofer* and then the secondary ones were added by the *naqdan*. The technology of printing has reduced Bible production into a single-step process.

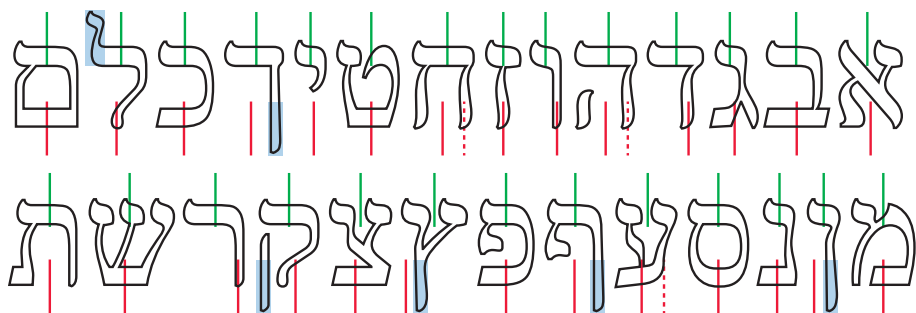


FIG. 5. Vertical axes of the graphs of the Monotype Peninim font. In green: upper vertical axes; in red: lower vertical axes; in dotted red: axes for placement of *patah furtivum*; in light blue: ascenders and descenders

3. Graphetic Challenges

The rules given in the previous section are extremely challenging on the graphetic level: it is often impossible to respect the perfect centering of diacritics with respect to vertical axes, and therefore, *niqqud* and *te'amim* may need to be decentered due to the graphetic context.

3.1. An Example

To illustrate the various Masoretic graphetic challenges, we use the example of word וַיֵּרָא ("and (he) saw" as in "And God saw the light, that it was good") from Gen 1:4.

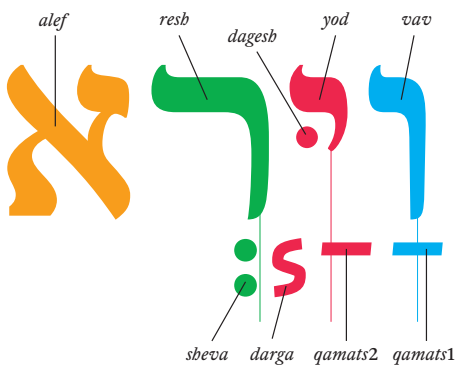


FIG. 6. The word וַיֵּרָא, analyzed

In Fig. 6 we illustrate the graphemic structure of this word: it has four letters: *vav*, *yod*, *resh*, and *alef*, and therefore four graphemic clusters. The *yod* carries a *dagesh*. As for sublinear secondary graphemes, the *vav* carries a *qamats* (we denote it by *qamats1*); the *yod* carries both a *qamats* (= *qamats2*) and the Masoretic diacritic *darga*; the *resh* carries a *sheva*.

The graphetic challenge lies in the fact that, on the one hand, the *yod* is a quite narrow letter so that the group of two secondary graphemes it carries (*qamats2* and *darga*) is necessarily wider than the letter itself, while, on the other hand, both letters surrounding the *yod* also carry sublinear secondary graphemes. The accumulated width of the four marks will fatally exceed the space assigned to them, which is delimited by the sublinear vertical axes of the *vav* and the *resh*.

In Fig. 6, typeset in the Monotype Peninim font by our system (Tiqwah v2), we have drawn the sublinear vertical axes of the three letters. We see that *qamats1* is perfectly centered on the vertical axis of the *vav* and that the *sheva* is slightly on the left of the axis of letter *resh*. The other noticeable fact of Fig. 6 is that the letters have kept their standard spacing, i.e., they are located at exactly the same distance as they would if they did not carry secondary graphemes.

Let us now take an overview, in Fig. 7, of how this word appears in 26 printed Bibles dating from 1517 to 2021 (#27 is our Tiqwah v2 output). The bibliographical data of each of the 26 Bibles is given in Appendix A (p. 320). 23% of these examples leave extra space on the left of the *yod*, 8% on its right, and 19% on both sides of the *yod*, i.e., half of these examples introduce parasitic spacing between letters and therefore break Rule #10. In many examples, the *darga*, which belongs to the *yod*, is placed on the axis of the *resh* and sometimes even to its left, introducing ambiguity into the visual recognition of graphemic clusters by raising the suspicion that the *darga* belongs to the *resh* since it is placed on the left of its vertical axis.

Example #1 contains only *niqqud* (it comes from the *Complutensian Polyglot Bible* (1517), in which the Hebrew text has only a very limited set of *te'amim*). Nevertheless, despite the fact that there is no *darga* in #1, the *sheva* has been moved to the left.

Example #2 is very important because it stems from the first printed Bible with full Masoretic marks. It was printed in 1525, in Venice, by the Antwerpian printer Daniel Bomberg. It is astounding that the word is typeset perfectly in #2, without any extra spacing between letters, and with the *sheva* perfectly centered on the *resh*'s axis. Among our examples, we have to wait until #22 (1999) to obtain a version of similar quality.

Example #3 moves the *qamats1* definitely too far to the right, and not at the same height as *qamats2*, but it otherwise manages to keep the *sheva* almost centered under the *resh*.

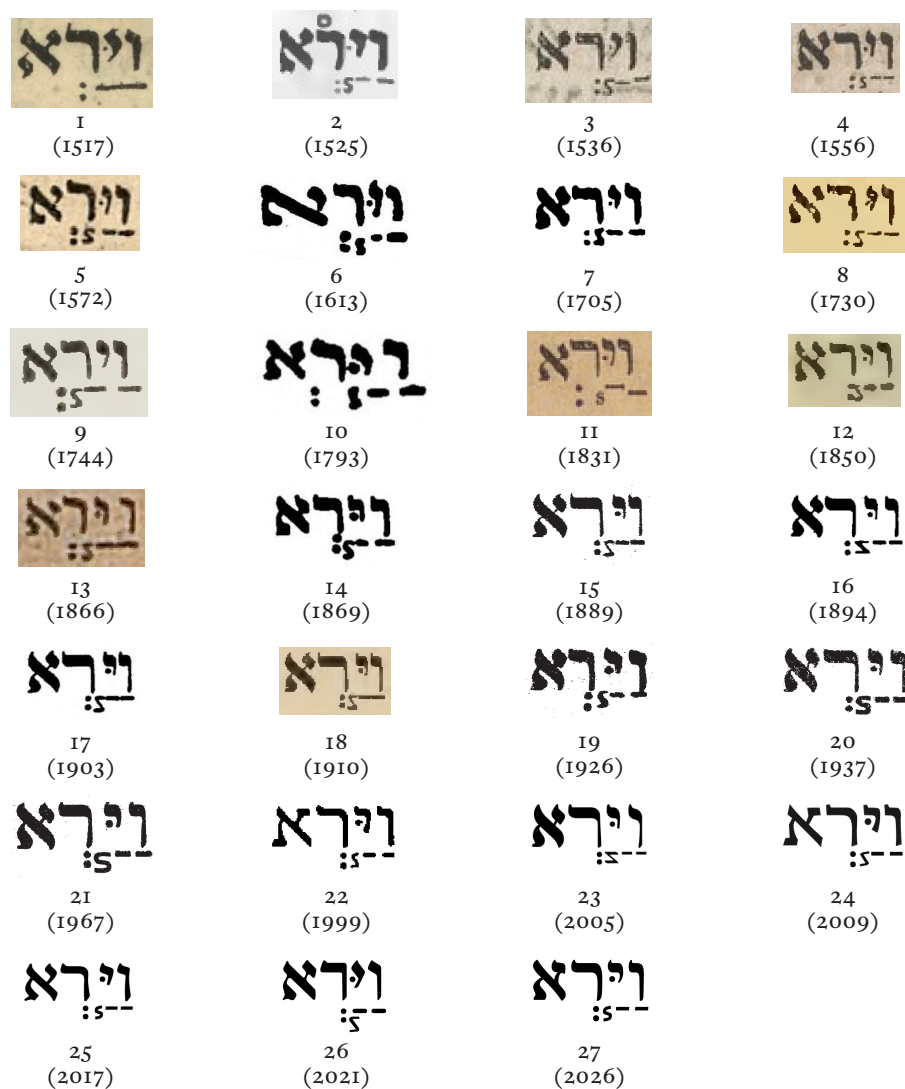


FIG. 7. The word **וירא** in 26 different printed Bibles, from 1517 to 2021 (#1–#26) and typeset by the Tiqwah v2 system in the Monotype Peninim font (#27)

Example #4 is quite good, but the *sheva* is slightly to the left, and it is not visually clear whether the *darga* belongs to the *resh* or to the *yod*. We encounter the same problem in #8.

Examples #5, #6, and #7 are of very bad quality and definitely ambiguous: the *darga* is centered on the axis of the *resh*, and the *sheva* is pushed to the left.

Example #9 is, by far, the worst of all cases: the *qamats2* of the *yod* has moved under the *resh*, so that the *darga* is pushed to the left and the *sheva* almost touches the *alef*.

Example #10 is also very bad: while the two *qamats* and the *darga* could be considered as correct, the *sheva* is pushed to the left for no reason, and appears at the boundary between the (quite large) *resh* and the *alef*.

Example #11 is from a 19th-century French edition of the Hebrew Bible. The *qamats1* is placed before the *vav*, the *qamats2* is correctly centered under the *yod*, but on a different height, probably to distinguish it from *qamats1*. The *darga* looks like an “s” letter and is placed on the *resh*’s axis, but a bit lower than the *qamats2*, and the *sheva*, once again, is pushed to the extreme left of the *resh*.

Example #12 has no parasitic spaces and the niqqud are well centered, but the *darga* is so narrow that it becomes unreadable, and touches the *sheva*.

Example #13 is reminiscent of Examples #9–#11, except that it has a lot of parasitic space between the *vav* and the *yod*.

Example #14 is quite interesting: the marks are quite legible, the letters are sufficiently spaced so as not to give the impression of parasitic space, but the most exciting fact is that the two *qamats* have different widths: *qamats2* is almost 50% the width of *qamats1*, and therefore leaves sufficient space to the *darga*. The *sheva* is only slightly moved to the left of the vertical axis.

Examples between 1889 and 1967 (#15–#21), as well as #23 (typeset using the Tiqwah v1 system, cf. Haralambous, 1994) and #25, use extra parasitic space between letters to solve the problem, the most extreme being #21, which is the only one that strictly respects the alignment of the *sheva* with the *resh*’s axis, even though it has a huge *darga* graph.

Examples #22 and #24 are set in the Hadassah font, which is very well suited for Masoretic text: it has large letters and narrow diacritics, so that no extra spacing is needed. The alignment is perfect.

Example #26 is a very special example: it comes from the *Koren Tanakh* and uses the Koren typeface. This typeface has been especially designed for Masoretic text, and we see that the *darga* is placed *under* the *qamats2*. This transgresses five centuries of printed Masoretic tradition, but the result is very pleasing and would be impossible to achieve otherwise, since the Koren graphs are very narrow and close together. Somehow we have the feeling that in 1831 the French printer of #11 also wanted to place the *darga* under the *qamats2* but didn’t manage to do so—#11 and #26 are the only examples where the *darga* is lower than the *qamats2*.

It is interesting to note that the technique of vertically stacking secondary graphemes has already been used in manuscript Bibles. Here is our example word in the Leningrad Codex (cf. Fig. 1 to see it in context):



We observe that (1) primary graphemes are indeed placed at their regular distances, (2) taking advantage of the fact that the *yod* descends only to half-height, the *qamats2* is raised to leave space for the *darga*. Nevertheless, the *naqdan* did not manage to center the *sheva* under the *resh*'s vertical axis, even though there was room enough to do so.

3.2. Algorithms

Algorithms solve this problem (cf. Haralambous (1994)) only partially by using letter spacing in extreme situations. In modern typefaces, OpenType features (Haralambous, 2007) address the problem to some extent. Nevertheless, OpenType features are limited because they process glyphs in linear order: in a situation where we have Cv_1Cv_2 , and we need to move v_1 and/or v_2 , a kerning rule involving them can hardly be applied since the glyph C lies between them.

To adapt a legacy-format font to the many interactions arising between clusters, one would have to include entire clusters as glyphs in the font. But this again presupposes that all difficult configurations are known to the font designer and can be anticipated. And that is not possible since, even though the Masoretic corpus (Hebrew Bible and New Testament) is relatively well-delimited, one cannot anticipate all textual variations in scholarly editions.

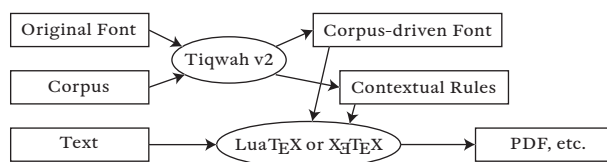
These considerations lead us to the conclusion that an optimal Masoretic font must be *built out of a given corpus*, and updated after every update or extension of the corpus, since every change may potentially introduce new problematic cluster interactions.

We call such a font a *corpus-driven* font. In the following section, we present our solution as well as the technologies involved.

4. The Tiqwah v2 System

4.1. System Architecture

The Tiqwah v2 system is an entirely rewritten version of the Tiqwah v1 system developed more than thirty years ago (Haralambous, 1994). Its architecture is as follows:



where the “Original Font” is any font containing standard Hebrew glyphs and *niqqud*, but not necessarily *te’amim*; “Corpus” is a corpus of texts using *te’amim* such as the Hebrew Bible, Margoliouth’s New Testament, etc.; “Text” is the text we wish to typeset (ideally the Hebrew text contained in “Text” should also be part of the “Corpus”); and “PDF, etc.” is the output. This system takes advantage of the separation between input and output in professional typesetting systems such as XeTeX or LuaTeX .

As for “Contextual Rules,” these are of two kinds: (1) reordering rules that ensure that glyphs are processed in a standard order that fits glyph names (see below) and (2) rewrite rules, with left and right context, which will map Unicode strings with lookahead- and lookbehind-endowed regular expressions, written either as a list of TECKit rules for XeTeX (Kew, 2021) or as a Lua program for LuaTeX (Hagen, Henkel, Hoekwater, and Scarso, 2025).

The output is searchable in any Acrobat-compliant PDF reader, since the glyphs of clusters have names that include the glyph names of all members of the cluster, in the logical order. So, for example, the cluster װ (which actually appears 282 times in the Bible, including in the first word of Genesis 1:1), the Unicode representation of which consists of characters U+FB2A (*shin* with *shin* dot), U+05B4 (*hiriq*), and U+0596 (*tipeba*), in this order, will have a glyph named uniFB2A05B40596, so that any search or copy action in the PDF file will use the underlying Unicode character string.

4.2. Volumetry

We adapted three fonts to Masoretic text: SBL Hebrew (designed by John Hudson in 2007), in which we kept the Masoretic glyphs but removed all OpenType features; Monotype Peninim (see Turkenich (2013) for the history of this emblematic font); Yarden (designed by Yarden Barzilai³ in 2023), a “playful, informal look [that] is great for light-hearted or humorous designs, comics, food packaging, and more”. We chose the Yarden font as an extreme Proof-of-Concept, because of its very free design. The reader can see in Figs. 9–11, on p. 322–324, the

3. <https://alefalefalef.co.il/font-yarden/>

beginning of the Book of Genesis typeset in Monotype Peninim, SBL Hebrew, and Yarden, as adapted by Tiqwah v2.

Let us give some elements of volumetry of the three fonts: SBL Hebrew, Monotype Peninim, and Yarden, all three built using the same corpus (Sefaria Tanakh⁴, Margoliouth’s New Testament⁵, and some other texts) of 160,099 distinct words containing at least one *niqqud* or *te’amim*:

	SBL Hebrew	Monotype Peninim	Yarden
# of added glyphs	9,625	12,877	18,664
# of .alt glyphs	2,108	5,111	10,898
avg. # of .alt	2.59	3.37	5.16
max # of .alt	101 (ך)	151 (ך)	123 (ך)
# of contextual rules	3,505	11,815	21,802
size of original *TF	61 kB	184 kB	30 kB
size of adapted OTF	449 kB	555 kB	755 kB
# of lines of .fea file	138	442	164
After font reduction with $\delta = 10$ glyph units (see §4.4):			
# of added glyphs	8,875	11,142	13,451
# of .alt glyphs	1,796	4,063	6,372
avg. # of .alt	1.72	2.32	2.82
max # of .alt	18 (ך)	27 (ך)	25 (ך)

where by “avg. # of .alt” we mean the average number of alternate glyphs for those that have at least one alternate glyph. We observe from the table that the number of glyphs of Yarden is almost double that of SBL Hebrew. This is due to irregularities in Yarden’s design, whereas SBL Hebrew has been tailored to be compatible with the Masoretic apparatus. The “.alt” glyphs are alternative versions of the same cluster, where one or more *niqqud* and *te’amim* have been moved for the cluster to be used in a specific context. The number of contextual rules depends on the number of alternative glyphs. As for the size of the OTF files obtained, it may seem small, e.g., when we compare them with CJK fonts, which also have tens of thousands of glyphs—this is because glyphs representing clusters are built using internal references rather than actual glyph code. The “.fea” file is the file of OpenType features, in Adobe syntax; its size represents the complexity of the font with respect to OpenType features.

In the following, we will describe the technology we use to build the cluster glyphs.

4. <https://www.sefaria.org/texts/Tanakh>

5. <https://www.jer-31.com/Margoliouth.html>

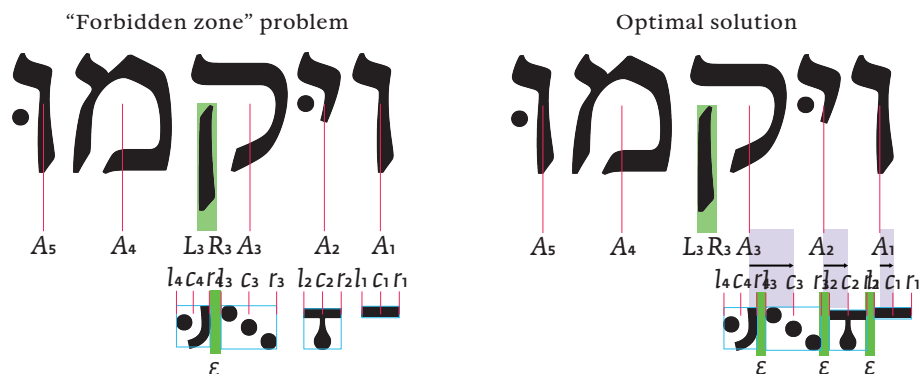


FIG. 8. The linear constraint minimization method applied to glyph positioning

4.3. Linear Constraint Minimization

Positioning of glyphs requires calculations in glyph space. In Fig. 8 we see the word **ויקמו** (Gen 22:19), in which we have drawn in red color the sublinear vertical axes, whose x-coordinates are called $A_1, \dots, A_5$. We have also drawn the four sublinear diacritics (three *niqqud* and one *te'amim*), each one carries three x-coordinates: the right boundary of its bounding box $r_1, \dots, r_4$, its stacking axis $c_1, \dots, c_4$ (which, depending on the diacritic's shape, are not necessarily the centers of the bounding boxes), and the left boundary of its bounding box $l_1, \dots, l_4$. We have drawn two zones in green color: one is the minimal distance ε between two sublinear (or supralinear) marks belonging to the same cluster, and the other, delimited by L_3 and R_3 , is a “forbidden zone,” namely the bounding box of the descender of the *qof*.

On the left part of the figure we have placed the diacritics according to their natural position *in the absence of “forbidden zone”*, that is we have matched the vertical axes c_1, c_2, c_3 of the three *niqqud* with the corresponding vertical axes of letters A_1, A_2, A_3 , and we have concatenated the *te'amim* next to the *niqqud* of the third letter, so that the distance between r_4 and l_3 is equal to ε . The problem is that this contradicts graphotactic Rule 9 (§ 2.4), namely that marks should not cross letter descenders, which, in our case, can be expressed as the constraint $l_4 - R_3 \geq \varepsilon$.

On the right part of the figure, we see the optimal solution: all diacritics are moved to the right so that $l_4 - R_3$, $l_3 - r_4$, $l_2 - r_3$, and $l_1 - r_2$ are all equal to ε . In other words, the presence of the *qof* has influenced the position of the diacritics of both letters preceding it. We have drawn the shift vectors of each diacritic in violet.

Let us model this situation in mathematical terms, keeping the same notation.

A *linear constraint minimization problem* consists of:

1. a linear function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ to be minimized by some value x ,
2. for which we have inequalities $g_i(x) \geq c_i$ for $i = 1, \dots, n$, where the g_i 's are linear functions $\mathbb{R} \rightarrow \mathbb{R}$.

The application of this mathematical problem to our situation is obvious: the linear function f is the sum of the absolute values of the differences between the c_i 's which respect graphotactic Rule 3 (left part of Figure 8) and those actually used; the g_i 's are the constraints given by graphotactic Rule 9, for example the fact that diacritics may not cross descenders (and even better: that they must stay at a distance of $\varepsilon + \kappa$, where κ is a potential kerning value. Other constraints occur at boundaries: a diacritic may not be “too far” on the right of the first letter of the word or on the left of the last letter of the word, etc.

This method provides us with optimal solutions in any situation. It will not work when the arrangement is physically impossible (this occurs if we have two descenders and the total width of the diacritics placed between them exceeds the width of the letter glyphs). These are extreme cases, and extreme cases require extreme solutions: we will either use narrower glyphs for the diacritics or stack them vertically. There is always a solution since these problems have been faced already and (mostly) solved by printers of the last five centuries...

4.4. Font Reduction

The problem with linear constraint minimization is that it provides optimal solutions up to a single glyph unit (which, for 10-point type, corresponds roughly to $3.53 \mu\text{m}$, far below the visibility threshold that, according to Westheimer (1981), is $\sim 10 \mu\text{m}$ at normal reading distance). Having distinct glyphs for clusters that differ by only a few glyph units is unnecessary and artificially conflates the font.

But how do we choose between two glyph clusters that differ by only a few glyph units, knowing that a cluster may have many diacritic components that may move in different directions (both horizontally and vertically)?

This is a typical graph theory problem, namely a *minimal clique partition in unit disk graphs*. Let S be a set of points in a Euclidean space $\mathbb{R}^d$, corresponding to the coordinates of the various alternate glyphs of a given cluster with d diacritics. Let us define a distance between members c, c' of S by $\text{dist}(c, c') := \sum |d_i - d'_i|$ where d_i, d'_i are the centers of the diacritics of c, c' .

Let us define the “ δ -near-graph” G as a graph the vertices of which are the elements of S and edges connecting vertices whenever their distance is less than δ . The “ δ -far-graph” G' is the complementary graph of G (same vertices but edges only where G has no edges). Covering S with

balls of diameter δ corresponds to finding cliques in G . But cliques in G correspond to discrete subgraphs in G' , and discrete subgraphs can be obtained by applying a coloring algorithm, and that is what we do.

Once we have found a minimal number of δ -balls covering S , we choose representatives for each, which are the only glyphs kept in the font.

As can be seen in the table of p. 316, font reduction allowed us to reduce the number of added glyphs by up to 28% (in the case of font Yarden). The impact of this change is most noticeable in the maximum number of .alts for a given glyph: it drops from 151 for Monotype Peninim to a mere 27.

5. Conclusion

We have argued that the Masoretic apparatus is best understood as a layered writing system in which a small set of primary graphemes carries a dense and tightly constrained set of secondary graphemes, and that the difficulty of typesetting it is not orthographic but graphetic. The marks themselves are well defined; the problem is realizing them on the page—centered on the correct axes, evenly spaced, clear of descenders and ascenders, and without disturbing the standard spacing of the consonants. This is a constraint-satisfaction problem that metal type, OpenType features, and even careful manuscript pointing solve only partially. The survey in Fig. 7 makes the point concretely: across five centuries, only a few editions place the marks of a single ordinary word without ambiguity or parasitic spacing.

Our response is the notion of a *corpus-driven font*. Because the difficult configurations arise from interactions between clusters, and because these cannot all be anticipated by a font designer, the font must be generated from the actual corpus to be typeset and regenerated whenever that corpus changes. Tiqwah v2 does this in two stages. Linear-constraint minimization computes, for every cluster, diacritic positions that respect the graphotactic rules and the “forbidden zones” of neighboring descenders, falling back on narrower marks or vertical stacking only in the rare physically impossible cases. A graph-coloring procedure then collapses clusters whose diacritics differ by less than the perceptual threshold into a single representative glyph, keeping the glyph count tractable even for a corpus of some 160,000 distinct pointed words. The output is a standard OpenType font whose cluster glyphs are named after their Unicode components, so the resulting PDF remains fully searchable.

That the same pipeline adapts fonts as different as SBL Hebrew, Monotype Peninim, and the playful Yarden—the last chosen precisely as an adversarial case—suggests the approach is general rather than tuned to one typeface, and it extends naturally beyond the Tanakh, as our

treatment of Margoliouth's cantillated New Testament shows. Several directions remain open: enlarging and diversifying the corpus (scholarly editions with non-standard pointing, the upper and lower systems of the Decalogue, Babylonian and Palestinian vocalization), refining the perceptual model that governs glyph merging, and applying the same corpus-driven, constraint-based methodology to other scripts whose diacritics interact as richly as the Masoretic ones—a portability the layered analysis of §2 was meant to support.

Acknowledgments

The author would like to thank Yosef Ginsberg and Gabriel Wasserman for their invaluable assistance in understanding Masoretic graphemics and graphetics and their guidance in improving the Tiqwah v2 system.

A. List of Bibles Used in Fig. 7

- #1 (Complutensian Polyglot Bible) *Vetus testamentu multiplici lingua nuc primo impressum. Et imprimis Pentateuchus Hebraico Greco atq[ue] Chaldaico idiomate, ediderunt: Antonius Nebrissensis, Demetrius Ducas, et al. Academia Complutensi Industria Arnaldi Guilielmi de Brocario: sumptibus F. Francisci Ximenez de Cisneros, Alcalá de Henares, 1514-1517*
- 2 (The Second Rabbinic Bible, first printed Hebrew Bible with the Massorah) *Great Scriptures with Onkelos' translation and commentary by Rasbi and Ibn Ezra, Venice, 1525* **מקרא גדולות עם תרגום אונקלוס ופירוש רש"י ואבן עזרא**
- 3 *Primus liber legis Mosaeicae, qui Genesis à Græcis inscribitur, ab Hebræis Beresith appellatur. Parisiis. Ex officina Christiani Wecheli sub scuto basiliensi, Paris, 1536*
- 4 *Liber Genesis. Parisiis. Ex officina Caroli Stephani (= Charles Étienne), Paris, 1556*
- 5 *Hebraicorum Biblicorum Veteris Testamenti Latina interpretatio. Christophorus Plantinus Regius Prototypographus Antuerpia excudebat, Antwerpen, 1572*
- 6 *Biblia Hebraica, cum interlineari interpretatione Latinæ Xantis Pagnini Lucensis, ex officina Plantiniana Raphelengii, Antwerpen, 1613*
- 7 *Biblia Hebraica, Secundum ultimam editionem Jos. Athias, a Jobanne Leusden denuo recognitam, Amstelædami et Ultrajecti, ediderunt Boom, Waesberge, etc., Amsterdam, 1705*
- 8 *Biblia Hebraica, ex aliquot manuscriptis et compluribus impressis codicibus, item massora tam edita, quam manuscripta, selectae variantes lectiones subiiciuntur cura ac studio Io. Heinr. Michaelis Halæ Magdeburgicæ, Typis & Jumptibus Orphanotrophei, Magdeburg, 1730*
- 9 *Book of Forty-Two, Main Section, Pentateuch, Mantua, 1744* **ספר ארבעה ועשרים החלק הראשון חמשה חומשי הורה כה מנטובה בשנת ויחכמ ידידיה שלמה לפ"ק**
- 10 *Biblia Hebraica olim a Christiano Reineccio edita et ad optimorum codicum et editionum fidem recensita et expressa nunc denuo ad fidem recensione masoreticæ ediderunt D. Io. Christoph. Doederlein et Ioannes Henricus Meisner, Lipsiæ, impensis*

- I. G. I. Breitkopfii et Soc.* (the same Breitkopf as in Breitkop & Härtel!), Leipzig, 1793.
- 11 *La Bible, traduction nouvelle, avec l'hébreu en regard, par Samuel Caben. Penteuque, La Genèse*, Paris, 1831
 - 12 *Biblia Hebraica, secundum ultimam editionem Jos. Athie, a Jobanne Leusden denuo recognitam. Ab Everardo van der Hooght, V.D.M. editio nova, recognita, et emendata, a Judab d'Allemand, Londini, typis excudebat A. Macintosh, impensis Jacobi Duncan*, London, 1850
 - 13 *A Torab scroll, Prophets, and Writings, accurately transcribed according to tradition, carefully studied by Yedidya ben Avram and the distinguished scholar Meir ba-Levi Letteris*, Berlin, 1866
ספר תורה נביאים ומכתובים מדויק היטיב יל פי המסורה
הוגה בעיון נמרץ על ידי ידידיה בן אברם ומאח החכם המובהק מהוי מאיר
הלוי לעטערס בערלין
 - 14 *Liber Genesis. Textum masoreticum accuratissime expressit, e fontibus masorae varie illustravit, notis criticis confirmavit S. Baer. Ex officina Bernhardi Tauchnitz*, Leipzig, 1869
 - 15 *Torab, Prophets, and Writings, accurate according to tradition, carefully studied by the distinguished scholar Meir ba-Levi Letteris*, Vienna, 1889
ספר תורה נביאים ומכתובים מדויק היטיב על פי המסורה הוגה בעיון נמרץ על ידי החכם המובהק מהו מאיר הלוי לעטערס וויען שנת ה'תרמ"ט ליצירה
 - 16 *Torab, Prophets, Writings, Masoretic Text according to Jacob ben Chayim and C.D. Ginsburg*, London 1894
 - 17 *The Holy Scriptures of the Old Testament, Hebrew and English, printed for the British & Foreign Bible Society*, Berlin, 1903
 - 18 *A Torab scroll, Prophets, and Writings, accurately transcribed according to the Masorab, carefully studied by the distinguished scholar Mordechai Meir ba-Levi Letteris*, New York: Hebrew Publishing Company, 1910
ספר תורה נביאים ומכתובים מדויק היטב על פי המסורה הוגה בעיון נמרץ על ידי החכם המובהק מהר מאיר הלוי לעטערס
 - 19 *Biblia Hebraica edidit Rud. Kittel*, Stuttgart, 1926
 - 20 *Biblia Hebraica edidit Rud. Kittel, textu masoreticum curavit P. Kable*, Stuttgart, 1937
 - 21 *Biblia Hebraica Stuttgartensia editio funditus renovata ediderunt K. Elliger et W. Rudolph, textum masoreticum curavit H. P. Rüger*, Stuttgart, 1967
 - 22 *JPS Hebrew-English Tanakh*, 2nd ed., The Jewish Publication Society, Philadelphia, 1999
 - 23 *Torab, Prophets, Writings, Masoretic Text according to Jacob ben Chayim and C.D. Ginsburg*, London 2005 (using the Tiqwah v1 system)
 - 24 *Hebrew-English Tanakh*, Varda Books, Skokie, Illinois, 2009
 - 25 Schmidt, Muhammad Wolfgang G. A., *The Complete Original Hebrew Text of the Old Testament or Tanakh. According to the Leningrad Codex*, Hamburg, disserta Verlag, 2017
 - 26 *The Magerman Edition. The Koren Tanakh, Hebrew-English*, Koren Publishers Jerusalem, 2021

1 בְּרֵאשִׁית בָּרָא אֱלֹהִים אֶת הַשָּׁמַיִם וְאֶת
 הָאָרֶץ: 2 וְהָאָרֶץ הִיְתָה תֵהוֹ וּבְהוּ וְחֹשֶׁךְ
 עַל-פְּנֵי תְהוֹם וְרוּחַ אֱלֹהִים מְרַחֶפֶת עַל-פְּנֵי
 הַמַּיִם: 3 וַיֹּאמֶר אֱלֹהִים יְהי אֹר וַיְהי-אֹר:
 4 וַיֵּרָא אֱלֹהִים אֶת-הָאֹר כִּי-טוֹב וַיְבַדֵּל
 אֱלֹהִים בֵּין הָאֹר וּבֵין הַחֹשֶׁךְ: 5 וַיִּקְרָא אֱלֹהִים
 | לְאֹר יוֹם וְלַחֹשֶׁךְ קָרָא לַיְלָה וַיְהי-עֶרֶב
 וַיְהי-בֹקֶר יוֹם אֶחָד: 6 וַיֹּאמֶר אֱלֹהִים יְהי
 רָקִיעַ בְּתוֹךְ הַמַּיִם וַיְהי מַבְדִּיל בֵּין מַיִם לַמַּיִם:
 7 וַיַּעַשׂ אֱלֹהִים אֶת-הָרָקִיעַ וַיְבַדֵּל בֵּין הַמַּיִם
 אֲשֶׁר מִתַּחַת לָרָקִיעַ וּבֵין הַמַּיִם אֲשֶׁר מֵעַל
 לָרָקִיעַ וַיְהי-כֵן: 8 וַיִּקְרָא אֱלֹהִים לָרָקִיעַ שָׁמַיִם
 וַיְהי-עֶרֶב וַיְהי-בֹקֶר יוֹם שֵׁנִי: 9 וַיֹּאמֶר
 אֱלֹהִים יִקְווּ הַמַּיִם מִתַּחַת הַשָּׁמַיִם אֶל-מָקוֹם
 אֶחָד וְתִרְאֶה הַיַּבְשָׁה וַיְהי-כֵן: 10 וַיִּקְרָא
 אֱלֹהִים | לַיַּבְשָׁה אָרֶץ וּלַמְקוֹהַּ הַמַּיִם קָרָא
 יַמִּים וַיֵּרָא אֱלֹהִים כִּי-טוֹב: 11 וַיֹּאמֶר אֱלֹהִים
 תְּדַשָּׂא הָאָרֶץ דָּשָׂא עֵשָׂב מִזְרִיעַ זֶרַע עֵץ פְּרִי
 עֹשֶׂה פְרִי לְמִינּוֹ אֲשֶׁר זֶרַע-בּוֹ עַל-הָאָרֶץ
 וַיְהי-כֵן: 12 וַתּוּצֵא הָאָרֶץ דָּשָׂא עֵשָׂב מִזְרִיעַ
 זֶרַע לְמִינֵהוּ וְעֵץ עֹשֶׂה-פְּרִי אֲשֶׁר זֶרַע-בּוֹ
 לְמִינֵהוּ וַיֵּרָא אֱלֹהִים כִּי-טוֹב: 13 וַיְהי-עֶרֶב
 וַיְהי-בֹקֶר יוֹם שְׁלִישִׁי: 14 וַיֹּאמֶר אֱלֹהִים

FIG. 9. The beginning of the Book of Genesis typeset in the Monotype Peninim font, adapted by Tiqawah v2

1 בְּרֵאשִׁית בָּרָא אֱלֹהִים אֶת הַשָּׁמַיִם וְאֶת הָאָרֶץ:
 2 וְהָאָרֶץ הָיְתָה תֹהוֹ וָבֹהוּ וְחָשֶׁךְ עַל־פְּנֵי תְהוֹם וְרוּחַ
 אֱלֹהִים מְרַחֶפֶת עַל־פְּנֵי הַמַּיִם: 3 וַיֹּאמֶר אֱלֹהִים יְהִי
 אֹר וַיְהִי־אֹר: 4 וַיֵּרָא אֱלֹהִים אֶת־הָאֹר כִּי־טוֹב וַיְבָרֶךְ
 אֱלֹהִים בֵּין הָאֹר וּבֵין הַחֹשֶׁךְ: 5 וַיִּקְרָא אֱלֹהִים | לָאֹר
 יוֹם וְלַחֹשֶׁךְ קָרָא לַיְלָה וַיְהִי־עֶרֶב וַיְהִי־בֹקֶר יוֹם אֶחָד:
 פ 6 וַיֹּאמֶר אֱלֹהִים יְהִי רָקִיעַ בְּתוֹךְ הַמַּיִם וַיְהִי מִבְדִּיל
 בֵּין מַיִם לַמַּיִם: 7 וַיַּעַשׂ אֱלֹהִים אֶת־הָרָקִיעַ וַיְבָרֶךְ
 הַמַּיִם אֲשֶׁר מִתַּחַת לָרָקִיעַ וּבֵין הַמַּיִם אֲשֶׁר מֵעַל לָרָקִיעַ
 וַיְהִי־כֵן: 8 וַיִּקְרָא אֱלֹהִים לָרָקִיעַ שָׁמַיִם וַיְהִי־עֶרֶב וַיְהִי־בֹקֶר
 יוֹם שֵׁנִי: פ 9 וַיֹּאמֶר אֱלֹהִים יִקְווּ הַמַּיִם מִתַּחַת הַשָּׁמַיִם
 אֶל־מָקוֹם אֶחָד וְתִרְאֶה הַיַּבְשָׁה וַיְהִי־כֵן: 10 וַיִּקְרָא אֱלֹהִים
 | לַיַּבְשָׁה אֶרֶץ וּלַמְּקוֹהַּ הַמַּיִם קָרָא יָמִים וַיֵּרָא אֱלֹהִים
 כִּי־טוֹב: 11 וַיֹּאמֶר אֱלֹהִים תְּדַשָּׂא הָאָרֶץ דָּשָׂא עֵשֶׂב מִזֶּרַע
 זֶרַע עֵץ פָּרִי עֵשֶׂה פְרִי לַמִּינֹה אֲשֶׁר זֶרְעוּבוּ עַל־הָאָרֶץ
 וַיְהִי־כֵן: 12 וַתּוֹצֵא הָאָרֶץ דָּשָׂא עֵשֶׂב מִזֶּרַע זֶרַע לַמִּינֹהוּ
 וַעֲץ עֵשֶׂה־פְרִי אֲשֶׁר זֶרְעוּבוּ לַמִּינֹהוּ וַיֵּרָא אֱלֹהִים כִּי־טוֹב:
 13 וַיְהִי־עֶרֶב וַיְהִי־בֹקֶר יוֹם שְׁלִישִׁי: פ 14 וַיֹּאמֶר אֱלֹהִים
 יְהִי מְאֹרֶת בְּרָקִיעַ הַשָּׁמַיִם לְהַבְדִּיל בֵּין הַיּוֹם וּבֵין הַלַּיְלָה
 וְהָיוּ לְאֹתוֹת וּלְמוֹעֲדִים וְלַיִּמִּים וְשָׁנִים: 15 וְהָיוּ לְמַאֲרֹת
 בְּרָקִיעַ הַשָּׁמַיִם לְהָאִיר עַל־הָאָרֶץ וַיְהִי־כֵן: 16 וַיַּעַשׂ אֱלֹהִים
 אֶת־שְׁנֵי הַמַּאֲרֹת הַגְּדֹלִים אֶת־הַמָּאֹר הַגָּדֹל לְמַמְשֶׁלֶת
 הַיּוֹם וְאֶת־הַמָּאֹר הַקָּטָן לְמַמְשֶׁלֶת הַלַּיְלָה וְאֵת הַכּוֹכָבִים:
 17 וַיִּתֵּן אֹתָם אֱלֹהִים בְּרָקִיעַ הַשָּׁמַיִם לְהָאִיר עַל־הָאָרֶץ:
 18 וּלְמַשֵּׁל בַּיּוֹם וּבַלַּיְלָה וּלְהַבְדִּיל בֵּין הָאֹר וּבֵין הַחֹשֶׁךְ
 וַיֵּרָא אֱלֹהִים כִּי־טוֹב: 19 וַיְהִי־עֶרֶב וַיְהִי־בֹקֶר יוֹם רְבִיעִי:
 פ 20 וַיֹּאמֶר אֱלֹהִים יִשְׂרְצוּ הַמַּיִם שָׂרָץ נֶפֶשׁ חַיָּה וְעוֹף

FIG. 10. The beginning of the Book of Genesis typeset in the SBL Hebrew font, adapted by Tiqwah v2

1 בְּרֵאשִׁית בָּרָא אֱלֹהִים אֶת הַשָּׁמַיִם וְאֶת הָאָרֶץ:
 2 וְהָאָרֶץ הִיְתָה תֹהוֹ וָבֹהוּ וְחָשֶׁךְ עַל־פְּנֵי תְהוֹם וְרוּחַ
 אֱלֹהִים מְרַחֶפֶת עַל־פְּנֵי הַמַּיִם: 3 וַיֹּאמֶר אֱלֹהִים יְהִי
 אֹר וַיְהִי־אֹר: 4 וַיֵּרָא אֱלֹהִים אֶת־הָאֹר כִּי־טוֹב וַיַּבְדֵּל
 אֱלֹהִים בֵּין הָאֹר וּבֵין הַחָשֶׁךְ: 5 וַיִּקְרָא אֱלֹהִים א
 לְאֹר יוֹם וְלַחָשֶׁךְ קָרָא לַיְלָה וַיְהִי־עֶרֶב וַיְהִי־בֹקֶר יוֹם
 אֶחָד: 6 וַיֹּאמֶר אֱלֹהִים יְהִי רָקִיעַ בְּתוֹךְ הַמַּיִם וַיְהִי
 מַבְדִּיל בֵּין מַיִם לַמַּיִם: 7 וַיַּעַשׂ אֱלֹהִים אֶת־הָרָקִיעַ
 וַיַּבְדֵּל בֵּין הַמַּיִם אֲשֶׁר מִתַּחַת לָרָקִיעַ וּבֵין הַמַּיִם אֲשֶׁר
 מֵעַל לָרָקִיעַ וַיְהִי־כֵן: 8 וַיִּקְרָא אֱלֹהִים לָרָקִיעַ שָׁמַיִם
 וַיְהִי־עֶרֶב וַיְהִי־בֹקֶר יוֹם שֵׁנִי: 9 וַיֹּאמֶר אֱלֹהִים יִקְוּ
 הַמַּיִם מִתַּחַת הַשָּׁמַיִם אֶל־מָקוֹם אֶחָד וְתִרְאָה הַיַּבְשָׁה
 וַיְהִי־כֵן: 10 וַיִּקְרָא אֱלֹהִים א לַיַּבְשָׁה אֶרֶץ וּלְמִקְוֵה
 הַמַּיִם קָרָא יַמִּים וַיֵּרָא אֱלֹהִים כִּי־טוֹב: 11 וַיֹּאמֶר
 אֱלֹהִים תִּדְשֵׁא הָאָרֶץ דָּשָׁא עֵשֶׂב מִזֵּרִיעַ זֶרַע עֵץ פֶּרִי
 עֹשֶׂה פֶרִי לְמִינֹו אֲשֶׁר זֶרְעוֹ־בּוֹ עַל־הָאָרֶץ וַיְהִי־כֵן:
 12 וַתּוֹצֵא הָאָרֶץ דָּשָׁא עֵשֶׂב מִזֵּרִיעַ זֶרַע לְמִינֵהוּ וְעֵץ
 עֹשֶׂה־פֶרִי אֲשֶׁר זֶרְעוֹ־בּוֹ לְמִינֵהוּ וַיֵּרָא אֱלֹהִים כִּי־טוֹב:
 13 וַיְהִי־עֶרֶב וַיְהִי־בֹקֶר יוֹם שְׁלִישִׁי: 14 וַיֹּאמֶר אֱלֹהִים
 יְהִי מְאֹרֶת בְּרָקִיעַ הַשָּׁמַיִם לְהַבְדִּיל בֵּין הַיּוֹם וּבֵין
 הַלַּיְלָה וְהָיוּ לְאֹתוֹת וּלְמוֹעֲדִים וּלְיָמִים וּשְׁנָיִם: 15 וְהָיוּ
 לְמֹאֲרֹת בְּרָקִיעַ הַשָּׁמַיִם לְהַאֲרִי עַל־הָאָרֶץ וַיְהִי־כֵן:
 16 וַיַּעַשׂ אֱלֹהִים אֶת־שְׁנֵי הַמְּאֹרֹת הַגְּדֹלִים אֶת־הַמְּאֹר
 הַגָּדֹל לְמַחְשֵׁלֶת הַיּוֹם וְאֶת־הַמְּאֹר הַקָּטָן לְמַחְשֵׁלֶת
 הַלַּיְלָה וְאֵת הַכּוֹכָבִים: 17 וַיִּתֵּן אֹתָם אֱלֹהִים בְּרָקִיעַ
 הַשָּׁמַיִם לְהַאֲרִי עַל־הָאָרֶץ: 18 וּלְמִשְׁלַל בָּיוֹם וּבַלַּיְלָה
 וּלְהַבְדִּיל בֵּין הָאֹר וּבֵין הַחָשֶׁךְ וַיֵּרָא אֱלֹהִים כִּי־טוֹב:

FIG. 11. The beginning of the Book of Genesis typeset in the Yarden font, adapted by Tiqwah v2

References

- At'ayan, Robert and Vrej N. Nersessian (1999). *Armenian Neume System of Notation*. Routledge.
- Bauer, Hans (1921). "Über die Anordnung der Suren und über die geheimnisvollen Buchstaben im Qoran". In: *Zeitschrift der Deutschen Morgenländischen Gesellschaft* 75, pp. 1–20.
- Futato, Mark D. (2020). *Basics of Hebrew Accents*. Zondervan Academic.
- Ginsburg, Christian D. (1897). *Introduction to the Massoretico-Critical Edition of the Hebrew Bible*. London: Trinitarian Bible Society.
- Hagen, Hans et al. (2025). "LuaTeX Reference Manual". <http://mirrors.ctan.org/systems/doc/luatex/luatex.pdf>.
- Haralambous, Yannis (1994). "Tiqwah, a Typesetting System for Biblical Hebrew, Based on TeX". In: *Actes du Quatrième Colloque International Bible et informatique, matériel et matière*. Amsterdam, pp. 445–470.
- (2007). *Fonts & Encodings. From Advanced Typography to Unicode and Everything in Between*. Sebastopol, CA: O'Reilly, pp. xx+1018.
- Hedánek, Jiří (2017). "Phonology of Masoretic Hebrew II. Accents as Prosody". PhD thesis. Charles University, Czech Republic.
- Hoop, Raymond de (2008). "Stress and Syntax; Music and Meaning: The Purpose and Function of the Masoretic Accentuation System". In: *Journal of Northwest Semitic Languages* 34, pp. 99–121.
- Jacobson, Joshua R. (1995). "The Cantillation of the Decalogue". In: *The Journal of Synagogue Music* 24.1, pp. 12–30.
- Kelley, Page H., Daniel S. Mynatt, and Timothy G. Crawford (1998). *The Masorah of the Biblia Hebraica Stuttgartensia*. Grand Rapids, MI & Cambridge, UK: William B. Eerdmans Publishing Company.
- Kew, Jonathan (2021). "The TECKit Language. Mapping byte encodings to Unicode". https://software.sil.org/downloads/r/teckit/TECKit_Language.pdf.
- Levine, Joseph A. (1989). *Synagogue Song in America*. Crown Point, IN: White Cliffs Media Company.
- McClintock, John and James Strong (1870). "Hebrew Version of the New Testament". In: *Cyclopedia of Biblical, Theological and Ecclesiastical Literature*. New York: Harper & Brothers.
- Mitchell, David C. (2017). "The Origins of the Masoretic Cantillation". A paper read at the Society of Old Testament Studies Centenary Conference, Nottingham University. URL: https://dlwqtxts1xzle7.cloudfront.net/51712224/The_Origins_of_the_Masoretic_teamim-libre.pdf.
- Nguyen, Martin (2012). "Exegesis of the ḥurūf al-muqaṭṭa'a: Polyvalency in Sunnī Traditions of Qur'ānic Interpretation". In: *Journal of Qur'anic Studies* 14/2, pp. 1–28.
- Osborn, J.R. (2017). *Letters of Light. Arabic Script in Calligraphy, Print, and Digital Design*. Princeton University Press.

- Park, Sung Jin (2020). *The Fundamentals of Hebrew Accents: Divisions and Exegetical Roles beyond Syntax*. Cambridge: Cambridge University Press.
- Tov, Emanuel (1999). "Paratextual Elements in the Masoretic Manuscripts of the Bible Compared with the Qumran Evidence". In: *Antikes Judentum und frühes Christentum: Festschrift für Hartmut Stegemann zum 65. Geburtstag*. Ed. by Bernd Kollmann, Wolfgang Reinbold, and Annette Steudel. Vol. 97. Beihefte zur Zeitschrift für die neutestamentliche Wissenschaft. De Gruyter, pp. 73–83.
- Troelsgard, Christian (2011). *Byzantine Neumes: A New Introduction to the Middle Byzantine Musical Notation*. Museum Tusculanum Press.
- Turkenich, Liron Lavi (2013). *From Frank Rübl to Peninim. A story of people, type and technology*. Master of Arts Thesis, University of Reading.
- Westheimer, Gerald (1981). "Visual hyperacuity". In: ed. by Hansjochem Autrum et al. Vol. 1. *Progress in Sensory Physiology*. Springer, pp. 1–30.
- Wickes, William (1881). *A Treatise on the Accentuation of the Three So-Called Poetical Books of the Old Testament*. Oxford: Clarendon Press.
- (1887). *A Treatise on the Accentuation of the Twenty-One So-Called Prose Books of the Old Testament*. Oxford: Clarendon Press.
- Wu, Andi and Kirk Lowery (2006). "A Hebrew Tree Bank Based on Cantillation Marks". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Ed. by Nicoletta Calzolari et al. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/6_pdf.pdf.
- Ἀθανασόπουλος, Γεώργιος Δ. [Athanasopoulos, Georgios D.] (1950). *Θεωρία τῆς βυζαντινῆς μουσικῆς [Theory of Byzantine Music]*. Πάτραι [Patras]: Τούλας-Μαυρίκιος [Toulas-Mavrikios].
- Хаздан, Евгения В. [Khazdan, Evgeniya V.] (2020). "От масоретских помет к знакам кантилляции: динамика восприятия [From Masoretic Notes to Cantillation Marks: The Dynamics of Perception]". In: *Богослужебные практики и культовые искусства в современном мире [Liturgical Practices and Sacred Arts in the Modern World]*. Vol. 4. Краснодар [Krasnodar]: Российский фонд фундаментальных исследований [Russian Foundation for Basic Research], pp. 288–310.

Audio LLM Subtokens as Encapsulated “Knowledge”

The Case of Persian Subtoken Graphemic Representations in Whisper

Behnoosh Namdarzadeh & Nicolas Ballier

Abstract. Whisper is a widely used, open-access Large Language Model (LLM) trained using a multilingual paradigm. As such, it represents an important opportunity for researchers to study how multilingual LLMs function across languages. In this paper, we probe the grapholinguistic representations of Whisper for an under-resourced language with a Perso-Arabic script: Persian.

Persian, with its abjad writing system, has a consonant-heavy inventory of letters and does not indicate vowels, making it a good candidate to test an audio LLM such as Whisper (Maleki and Ahrenberg, 2008b). Large Language Models do not deal with tokens, morphemes or graphemes as linguists know them but subtokens, fragments of words generated by an algorithm. This subtokenisation process creates fracto-morphemes, clusters of letters, that continue to represent language (this is literally “natural language processing”) but the relationship of the letters to morphology, semantics and phonology is completely changed.

We want to investigate this technical paradigm shift for writing systems such as Persian, by analysing the sounds to letters mapping as mediated by acoustic representations and LLM subtokens. Using a C++ implementation (Gerganov, 2003) of Whisper (Radford et al., 2023), we have extracted the subtoken dictionary of this model and their probabilities using a customised C++ implementation of Whisper.

In this work, we probe the phonotactic and graphotactic status of the subtokens representing Persian in order to understand how “knowledge” (representation) is distributed in the Whisper subtokens used to transcribe Persian. Understanding the subtokenisation process of large language models is crucial for advancing both NLP techniques and linguistic theories. Investigating how these models represent languages such as Persian sheds light on the intricate relationship between the acoustic signal, subtoken structures, and linguistic properties.

Our experiments show the division of labour between the encoder and the decoder. We will show that the decoder is used to post-process the data, as evidenced by the fact that the internal representations of the Whisper models are only the isolated forms of graphemes and limited digraphs. We demonstrate that Whisper models tend to select the formal variant in their transcription. For example, in a read speech of 100 sentences, Whisper produces the formal variant of the object marker *rā* in the transcriptions three times more than its informal

Behnoosh Namdarzadeh¹  0009-0000-8730-7766,

Nicolas Ballier²  0000-0003-2179-1043

^{1,2}ALTAE, Université Paris Cité

¹E-mail: behnooshnamdar@gmail.com, ²E-mail: nicolas.ballier@u-paris.fr

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 327–360. <https://doi.org/10.36824/2024-graf-namd>
ISBN: 978-2-487055-10-0, e-ISBN: 978-2-487055-11-7

variant *ro*, even though the speaker only produces the informal variant in the read speech.

1. Introduction

Whisper is an automatic speech recognition (ASR) system developed using large-scale weak supervision. It was trained to predict transcripts of audio from 680,000 hours of labeled audio data, with 117,000 hours covering 96 languages other than English. The audio was broken into 30-second segments and paired with transcripts from different sources. Audio was resampled to 16,000 Hz, and an 80-channel log-magnitude Mel spectrogram was computed. Whisper was trained for several tasks such as language identification, transcription, translation, and time-stamps prediction. Whisper’s training included data for 96 languages besides English, including Persian, but with only 392 hours for the translation task and 24 hours for the transcription task. The performance for transcription of Persian is reported with a word error rate of 44.8 for the large model (compared to 6.3 for English), and the impact on performance of the amount of training data is consistent with the expectations (Ballier, Burin, et al., 2024). The training data transcripts were tokenised at subword level using a byte pair encoding algorithm (Sennrich, Haddow, and Birch, 2016) that was applied to the 96 languages of the training data, resulting in a multilingual vocabulary of units we call “subtokens” in the rest of the paper.

Whisper is a neural network that learns from a training corpus made up of audio recordings (690,000 hours for the largest model), combined with their transcription or translation. The aim is to match frames (windows of the sound signal) with (sub)tokens, i.e., arbitrary sequences of characters. It is therefore a supervised learning algorithm. Whisper is therefore an auto-regressive algorithm: given an audio input and the vector representation it constructs, it predicts the tokens one after the other to generate the final output. More specifically, based on the input audio and the prefix already generated (the previously predicted tokens), Whisper’s decoder determines a probability distribution for all the tokens in its vocabulary. By choosing the subtoken with the highest probability at each stage (or by sampling according to this distribution), it gradually builds the output sequence.

We wish to explore the graphophonemic status of these subtokens as applied to the transcription task (ASR) of Persian. These LLM subtokens are a paradoxical object that are reminiscent of splinters but that are not morphological entities. We want to elaborate the inventory of subtokens used by the multilingual Whisper for Persian transcription by investigating the mapping between the acoustic realisations of the Persian data used as input and the Whisper prediction. In this sense, we address the

presumed encapsulated “knowledge” in the subtoken dictionary of the speech model. An important aspect to bear in mind is the potential misleading metaphor that if LLMs have representations they “understand” speech or “recognise” words. Even though the main type of encoded representation is a dictionary of subtokens that serves as a repository for the corresponding numbers of vectors of numerical figures, that serves as a form of representation, if any, alternatively for every 30 second speech window. The different attention weights corresponding to the internal representations can also be extracted and potentially investigated (see Mohebbi, Chrupała, Zuidema, and Alishahi (2023) as an example). Linguists still need to unravel what becomes of the different languages that are processed using these multilingual models. In this respect, it is likely that the subtokenisation phase is not innocuous (Petrov, La Malfa, Torr, and Bibi, 2024). We partially discuss the phonology of the Whisper dictionary of subtokens¹, the subtokens used to represent Persian writing in the LLM. Persian offers a unique perspective for this type of investigation for three reasons. First, the scarcity of the training data (24 hours of training data in Whisper) entailed a scarcity of subtokens in the Whisper dictionary used to transcribe Persian. Second, we have a privileged window on the phonographematics of Whisper since the acoustic space is transcribed by a limited amount of subtokens, mostly Persian isolated characters, whereas Persian has graphotactic constraints for final, medial or initial letters. Third, Persian shares many characters with Arabic, which has more training data in Whisper. Therefore, Persian serves as a test case for the investigation of a language that shares some of the characters of the same Unicode group (Perso-Arabic script).

As an abjad, rather than an alphabet, the Persian written language does not note vowels; this frequently results in an error in the transcription of vowels. 28 letters of the Arabic script have been supplemented with 4 additional consonants in order to preserve consonantal distinctions that do not exist in Arabic. However, Arabic is also encoded in the Whisper dictionary of subtokens. Previous studies address the challenge of adapting multilingual language models to specific languages for different NLP tasks (Downey, Blevins, Goldfine, and Steinert-Threlkeld, 2023).

The rest of the paper is structured as follows: Section 2 presents the tenets of the Persian writing system and the variety of Persian we describe. Section 3 presents the phonological features of Persian and subtokenisation process in Whisper. Section 4 details our methodology and the datasets we used in our experiments. Section 5 presents our experiments with the probing of the subtoken dictionary and discusses the distribution of graphemes encapsulated in the Whisper subtokens used to transcribe Whisper. We analyse our results in Section 6 and Section 7 discusses them. Finally, we conclude in Section 8.

1. We do not discuss the special subtokens, see Ballier, Burin, et al. (2024).

2. Background on Persian Writing System

This section provides an overview of Persian graphematics and examines the graphotactic constraints that may influence the performance of audio-based LLMs.

The history of the Persian language is divided into three periods: Old Persian (c. 525 BC–300 BC), Middle Persian (c. 300 BC–800 AD) and Modern Persian (800 AD to the present). After the advent of Islam in 642 AD, Persians adapted the Arabic alphabet to develop the modern Persian alphabet. The Arabic alphabet has 28 letters, but Iranians added four additional letters <پ>, <چ>, <ژ>, <گ> to create the current 32 Persian letters (Rastorgueva and Hill, 1964). Modern Persian emerged in the 9th century after this adoption of the alphabet and the borrowing of many Arabic words.

The Persian script, technically called “abjad”, consists of 32 letters, 28 of which are shared with Arabic, while the remaining four are unique to Persian. Persian adopts a non-Latin script written from right to left (Baluch and Choudhury, 2011). It comprises six vowels: three short vowels, marked with diacritical marks, and three long vowels, marked by certain consonants (*ibid.*). Therefore, Persian writing can be called a “consonantal alphabet”. Diacritics for short vowels are mainly used for elementary readers, disappearing from textbooks after the first grade (Rahbari and Sénéchal, 2009), and are generally absent from standard written texts. Long vowels, in contrast, are consistently represented in writing.

The relationship between graphemes and phonemes in Persian is highly consistent. Each grapheme corresponds to a single phoneme (Arab-Moghaddam and Sénéchal, 2001). Therefore, Persian is classified as having a shallow orthography when short vowels are represented or when only long vowels and consonants are present (Baluch and Choudhury, 2011). However, the omission of short vowels results in a deep and opaque orthography where the relationships between graphemes and phonemes are inconsistent.

2.1. Graphotactic Constraints for Persian

This subsection explores the graphotactic constraints of Persian², which may affect the performance of LLMs such as Whisper (Maleki and Ahrenberg, 2008a).

2. It should be mentioned that the dialect we are examining for the sets of our experiments is a variant of Farsi, called “Tehrani”, the standard colloquial dialect spoken in Iran (Karimi, 2005). We use the term “Persian” to refer to this variant in this paper.

Ninety different forms, referred to as “allographs”, are needed to represent the 28 letters of the Persian alphabet. Persian characters can have different graphematic forms depending on their positions. The isolated form of characters (28 letters) depends on their position within a word. The isolated form of each letter of the 28 letters can vary depending on its contextual placement, taking on different forms (allographs) in the initial, medial, or final positions (ibid.). The training data may not include all the required cases to ensure an ideal distribution of these contextual character forms, specifically Segment-Initial, Segment-Medial, Segment-Final, and Isolated forms. The dictionary of Whisper subtokens (see Appendix, Table 5) will be comprehensively analysed in Section 5.2.

In Persian, a single phoneme can be represented by multiple letters, resulting in graphematic inconsistencies. This variation requires practice and familiarity to determine which specific grapheme is used in a given word. For instance, the phoneme /t/ is represented by two letters: <ت> and <ط>; the phoneme /h/ by two letters: <ه> and <ح>; the phoneme /s/ by three letters: <س>, <ص>, and <ث>; and finally, the phoneme /z/ by four letters: <ذ>, <ز>, <ض>, and <ظ>. This is particularly problematic when transliterating foreign words and selecting the appropriate grapheme for a given phoneme. For example, the graphematic representation of the French-origin word *billet* and the two variants of /t/ phoneme is controversial. This has resulted in two different graphematic forms: <بلیت> and <بیط>. Table 3 illustrates the various letters that correspond to the same phoneme. However, in Arabic, each of these letters corresponds to a distinct phoneme.

3. Phonological System of Persian and the Speech-to-Subtoken Mapping

We first recall the phonological system of Persian before explaining how sounds are mapped to the inventory of subtokens. In this paper, we follow the standard phonological convention that indicates graphemes (letters) with angled brackets (<>), realisations in square brackets and phonemes (or targets) with slanted bars (/).

3.1. Vowels and Consonants

The standard Persian of Iran has six vowel sounds as shown in Table 1. The vowels /a, e, o/ are short vowels and /i, u/ are long vowels. Since the Persian writing system is an abjad (consonantal alphabet), short vowels do not have specific letters and are not written in written Persian, and only long vowels appear in texts. Long vowels are always marked

with consonant letters. The short vowels, as mentioned earlier, are not marked in standard written Persian, but for learners of Persian, they can be precisely marked with diacritics. Inserting diacritics is not allowed in the early years of primary school.

The three long vowels, /ɒ, i, u/, are frequently found word-finally. Less commonly occurring word-finally are the short vowels, /a, e, o/. While /e/ is present in some high-frequency words, word-final /o/ is rare (still appears in some high frequency words.)

Any vowel may begin a word, though some occur more frequently than others. Word-initial /a/ is highly prevalent, while word-initial /i/ is less common. Word-initial /u/ is rare. In summary, lax vowels such as /e/ and /a/ are the most common in word-initial positions, while word-initial /o/ is seldom encountered.

TABLE 1. Tehrani Persian Vowel Chart

<i>Height</i>	<i>Front</i>	<i>Back</i>
<i>Close</i>	i	u
<i>Mid</i>	e	o
<i>Open</i>	a	

Due to the complexities of inserting or omitting vowels in written Persian, significant challenges arise in accurately transcribing vowels. There are some factors such as the neighbouring graphemes and the type of syllable containing the vowel which affect the choice of grapheme (allographs) for the vowel (Maleki and Ahrenberg, 2008a). Persian script often omits short vowels, as seen in /ketɒb/ (<کتاب>), where the short vowel /e/ is not represented, reflecting the inherent complexity of vowel representation in the language. Additionally, vowel representation is influenced by neighbouring graphemes, as in /χɒphar/ (<خواهر>) where the character <و> does not denote a /v/ or /u/ sound but forms part of a <خو> digraph for /χ/ followed by /ɒ/.

This subsection explores the constraints that govern the distribution and arrangement of consonants in written Persian. Persian includes 23 consonant sounds, each represented by a dedicated letter. The table of International Phonetic Alphabet (IPA) symbols for these consonants is presented in Table 2. Notably, the representation of a letter in Persian script varies depending on its position within a word. A single letter may assume different forms when appearing in initial, medial, or final position of a word [See Table 3]. This positional variability is a defining characteristic of the Persian writing system. Word-initial consonant clusters are absent in Persian, and Persian speakers tend to insert a vowel between consonants when pronouncing borrowed words with initial clusters from the source language, thus avoiding word-initial conso-

nant clusters. Alternatively, a vowel may be inserted before the cluster, leading to a VC+C sequence (Mahootian and Gebhardt, 1997). Moreover, word-initial clusters can be broken up, resulting in a CV+C sequence. Various types of two-consonant clusters are permissible in syllable-final positions (e.g., nasal+plosive: /omq/ (<عمق>) meaning ‘depth’).

TABLE 2. Persian Consonants Organized by Manner and Place of Articulation

	<i>Labial</i>	<i>Alv.</i>	<i>Post-alv.</i>	<i>Velar</i>	<i>Uvular</i>	<i>Glottal</i>
<i>Nasal</i>	m	n				
<i>Stop</i>	p, b	t, d	tʃ, dʒ	k, g	(q)	ʔ
<i>Fricative</i>	f, v	s, z	ʃ, ʒ	x, ɣ		h
<i>Tap</i>		ɾ				
<i>Approx</i>		l	j			

3.2. The Semiotics of Subtokenisation

In this subsection we explain how the Persian representation between graphemes and meaning is mediated with the subtokenisation phase (Sennrich, Haddow, and Birch, 2016). Subtokens are a hybrid representation that bear testimony to the frequency of tokens in the training data of the subtokenisation algorithm (and of Whisper’s training). Subword units have emerged as a crucial component in language modeling. They play a foundational role in the tokenisation processes of advanced transformer-based language models, such as those utilising Byte Pair Encoding (BPE) (ibid.). This subword-based tokenization method underpins the architecture of many LLMs (Manova, 2023). The adoption of subword tokenization strategies, like BPE, has significantly contributed to the success of pre-trained transformer models in natural language processing, enabling them to effectively handle a diverse array of languages and linguistic structures (Bostrom and Durrett, 2020). The (sub)tokens are generated automatically from the training data using tokenisation algorithms. These algorithms divide the text into units based on the frequency with which sequences appear in the corpus. The main algorithms used are Byte Pair Encoding (BPE) (Sennrich, Haddow, and Birch, 2016), WordPiece, initially developed for BERT (Devlin, Chang, Lee, and Toutanova, 2019), and SentencePiece (Kudo, 2018). Depending on their frequency of appearance, (sub)tokens can correspond to whole words, sequences of 2 or 3 characters, or even a single character.

Subtokenisation can be challenging for Persian due to its graphematic and phonographematic constraints, as outlined in Table 2. In this context, the implementation that predicts on a sub-token basis may

differ significantly from a representation based on entire sentences. Typically, isolated tokens are expected to correspond systematically to individual letters. Figure 1 illustrates the subtokenisation in three languages—French, English, and Persian—using the Whisper tokeniser³. As can be seen with these examples, the frequent tokens in English are less likely to be subdivided into subwords (subtokens) than French or Persian words.

FIG. 1. Subtokenisation in French, English, and Persian sentences using Whisper tokeniser

To investigate the frequency of subdivision for Persian words, we tested the Persian FLEURS dataset (Conneau et al., 2023) for a distributional analysis to observe the frequency of the subtokens for Persian and estimate the inventory of subtokens from the Whisper dictionary used to transcribe Persian data.

3.3. Hallucinations

Another aspect of the Whisper dictionary of subtokens used for transcriptions is the fact that the Whisper multilingual dictionary of subtokens includes all the subtokens used to transcribe the 99 languages with which Whisper was trained. With Whisper smaller models (tiny, base and small)⁴, hallucinations can be observed, subtokens belonging to other languages groups are used in the Whisper transcription. One way to analyse the evolution of hallucinations across the size of the models is to observe that the hallucinations in the outputs belong to different language groups, and, in that respect, they somehow follow the pecking order of the hierarchy of languages used in the training data. The Chinese characters are overwhelmingly present in the transcriptions of the tiny models and that is consistent with the role of the size of the Chinese training data.

3. https://huggingface.co/docs/transformers/model_doc/whisper

4. With 39M,74M and 244M parameters respectively (Radford et al., 2023).

4. Material and Methods

This section outlines the materials and methods used in the study.

4.1. Materials

Here are the three datasets used in our experiments:

- YouTube Audio File: A 69-minute audio recording extracted from YouTube⁵ was selected for analysis. The recording features three male speakers engaging in a discussion on social issues. Due to the spontaneous and unstructured nature of the conversation, manual transcription would be both challenging and time-consuming. By utilizing OpenAI’s Whisper for transcription, we gain two significant advantages: the generation of an accurate transcription and its accompanying translation. This output provides a valuable foundation for developing a bilingual corpus.
- Common Voice Mozilla: Additional Persian data for detailed experiments was sourced from the Common Voice Mozilla project⁶, a diverse dataset of speech recordings designed for speech recognition tasks.
- FLEURS Dataset: The FLEURS dataset⁷, which focuses on multilingual speech recognition and understanding, was also used to supplement our analysis and broaden the scope of our Persian corpus experiments.

4.2. Transcription Method With OpenAI’s Whisper

Our approach focused on leveraging a customized C++ implementation of Whisper⁸, building upon the foundational work by Radford and colleagues (Radford et al., 2023). Whisper offers six model variants, ranging from tiny to large-v1, which provide valuable insights into the mapping of spoken Persian to written Persian. The Whisper.cpp implementation by Gerganov⁹ facilitates efficient processing of Whisper’s parameters and includes tools for visualizing subtokens along with their associated probabilities. This functionality enables access to probability scores for each subtoken generated by Whisper, forming the basis for

5. <https://www.youtube.com/watch?v=YsMDE0EH42s>

6. <https://commonvoice.mozilla.org/fa>

7. <https://huggingface.co/datasets/google/fleurs>

8. <https://github.com/jbyunes>

9. <https://github.com/ggerganov/whisper.cpp>

developing a calibration model to evaluate the accuracy of the Whisper models.

4.3. Documenting Whisper’s Dictionary of Subtokens Used for Persian

We compared two methods. We first compiled the prediction at subtoken levels extracted for the different models. Then we ran the tokenisation algorithms on the Persian component of the FLEURS training data to compare the distribution of subtokens. we used a subsample of the FLEURS dataset comprising 85 hours of speech. The FLEURS dataset¹⁰ (Flores Enhanced Universal Representation for Speech) is a multilingual speech dataset, serving as the speech-focused counterpart to the FLoRes machine translation benchmark. It builds on the FLoRes dev and devtest datasets, comprising 2009 n-way parallel sentences translated into 102 languages. This extensive coverage makes FLEURS one of the most valuable resources for advancing multilingual speech research.

A key strength of FLEURS lies in its use of multilingual fine-tuning, which allows models to harness shared linguistic features across diverse languages. The dataset’s results are thoughtfully organised into seven geographical regions, showcasing the dataset’s cultural and linguistic breadth. Among the languages included is Persian, highlighting the dataset’s inclusion of languages from the Central Asia/Middle East/North Africa region.

5. Experiments

This section describes the experiments conducted in this study on written Persian. To better understand the distribution of graphemes in Persian, we used the FLEURS training dataset (Conneau et al., 2022), which is publicly available in 102 languages and published by Google. We analysed the distribution of Persian graphemes across this dataset. The dataset contains a total of 280,786 graphemes, with 34 unique graphemes. Figure 2 shows the distribution of graphemes in written Persian. We hypothesize that the more frequently a subtoken appears in publicly available online corpora, the fewer challenges LLMs like Whisper encounter in accurately processing it. Frequent exposure to common subtokens during training allows the model to develop a stronger representation and mapping for these elements, reducing errors in transcription and improving overall reliability. Conversely, subtokens that

10. <https://huggingface.co/datasets/google/fleurs>

are rare or less represented in the training data are more likely to lead to inaccuracies, reflecting gaps in the model’s learned patterns.

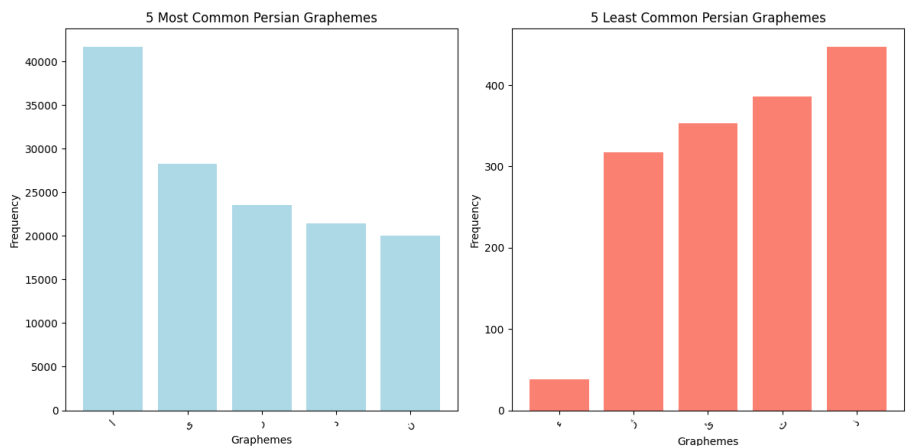


FIG. 2. Distribution of Persian Graphemes in the FLEURS dataset. As the different frequency scales indicate, the order of magnitude between most frequent and least frequent graphemes is about one hundred to one.

5.1. Distributional Frequencies in the Transcription of a Reference Dataset

To improve the performance of Natural Language Processing (NLP) tasks, the quantity of training data plays a crucial role (Ballier, Burin, et al., 2024). In general, an increase in the volume of training data directly correlates with improved task outcomes, as measured by various performance metrics. For instance, in Automatic Speech Recognition (ASR) tasks, an increase in the availability of training data for a particular language tends to lower the Word Error Rate (WER), indicating enhanced model accuracy and reliability. Figure 4 illustrates the distribution of graphemes in the FLEURS dataset.

The aim is to examine the frequency of consonants and vowels that appear at the beginning and end of words in written Persian, to be later compared to the analysis of subtokenisation outputs produced by Whisper. As shown in Figure 3, the characters <د>, <م>, <ب>, and <ک> typically appear at the beginning of words, while the consonants <ر>, <ذ>, <ه>, and <ن> typically appear at the end of words. Regarding vowels, the vowel <ا> frequently appears both at the beginning and at the end of

words. If we consider the FLEURS corpus as a foundational dataset for Persian, it is reasonable to assume that Whisper, when trained on similar data, has been exposed to similar distributions of graphemes. Specifically, Whisper may be more likely to accurately transcribe graphemes that are highly represented in its training data. Conversely, graphemes that are under-represented in the FLEURS dataset might pose challenges, potentially leading to transcription inaccuracies due to limited exposure during training.

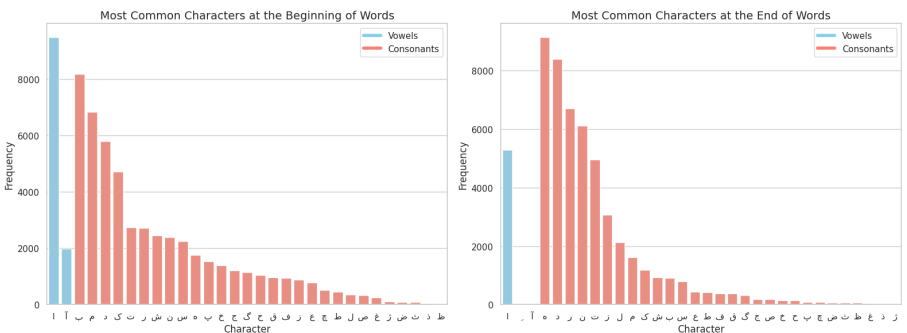


FIG. 3. Positional distribution of consonants and vowels at the beginning and end of words

5.2. Distributional Frequencies in Whisper Transcriptions of the FLEURS Speech Training Set

This subsection analyses Whisper transcription (large model) of the FLEURS speech training set. We then compare the inventory of Whisper subtokens used for transcriptions with the positional constraints of Persian graphemes.

Table 3 summarises the findings of this analysis. The first column presents the IPA representations of Persian characters. The second column lists the isolated forms of Persian graphemes, accompanied by a binary code (0–1) indicating their presence in Whisper’s subtoken dictionary. A value of 1 signifies that the grapheme is included in the dictionary, while a value of 0 indicates its absence. For the isolated representation of graphemes, it was observed that $\langle \text{ج} \rangle$, which belongs exclusively to the Persian alphabet and not the Arabic script, is not present in the Whisper multilingual dictionary. The fourth column of Table 3 shows the final graphematic representations of characters. Notably, the

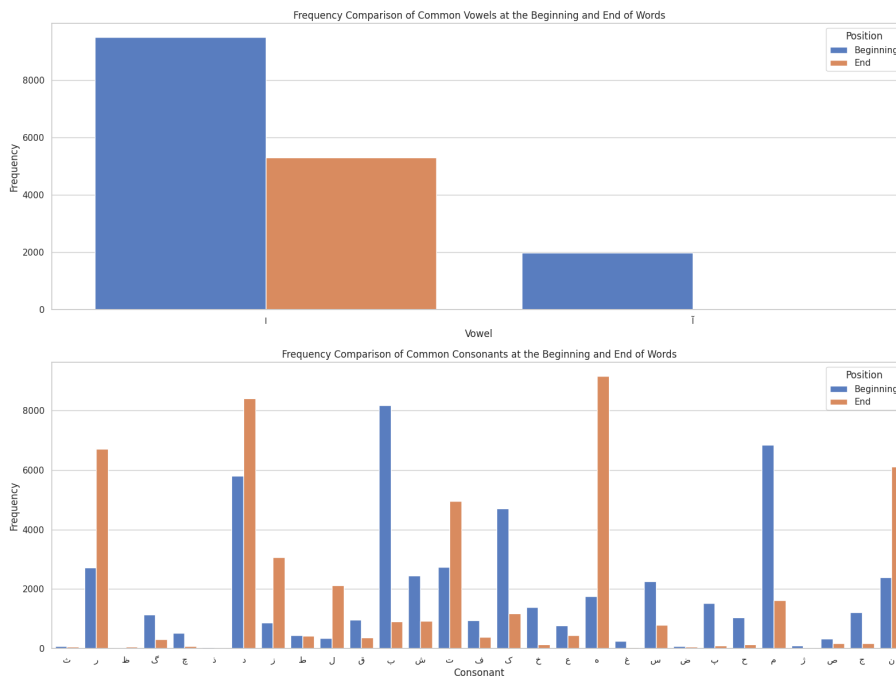


FIG. 4. Distribution of common consonants and vowels at the beginning and at the end of words in the Whisper transcriptions of the FLEURS speech training set (85 hours)

characters <گ, غ, ظ, ز, ژ, چ, پ> are absent from Whisper’s subtoken dictionary in their final forms. It is important to highlight that <پ ژ چ> are Persian-specific graphemes. The sixth column represents the medial forms of graphemes in Persian. In Whisper’s subtoken dictionary, we did not observe any Persian characters in their medial forms, except for <ی, ن, ل, ع, ت>. The eighth column reports the initial forms of characters within Persian tokens, with the following characters being absent in their initial forms: <گ, غ, ظ, ط, ض, چ, ج, پ>.

By accessing the dictionary of subtokens for all 99 languages supported by Whisper, we extracted the subtokens associated with the Perso-Arabic script. We identified four possible positions for Persian characters within a word: initial, medial, final, and isolated forms¹¹.

11. Some entries in the columns of graphematic representation for Persian contain a value of 0, indicating that these forms do not exist in the Persian script.

TABLE 3. Whisper Coding of Persian Letters in its Subtoken Dictionary

<i>IPA</i>	<i>Isolated form</i>	<i>Whisper isolated</i>	<i>Final form</i>	<i>Whisper final</i>	<i>Medial form</i>	<i>Whisper medial</i>	<i>Initial form</i>	<i>Whisper initial</i>
/v/	ـ	1	ـ	1	0	0	0	0
/b/	بـ	1	بـ	1	بـ	0	بـ	1
/p/	پـ	1	پـ	0	پـ	0	پـ	0
/t/	تـ	1	تـ	1	تـ	1	تـ	1
/s/	ثـ	1	ثـ	1	ثـ	0	ثـ	1
/d/	دـ	1	دـ	1	دـ	0	دـ	0
/t/	ذـ	1	ذـ	0	ذـ	0	ذـ	0
/h/	حـ	1	حـ	1	حـ	0	حـ	1
/χ/	خـ	1	خـ	1	خـ	0	خـ	1
/d/	ذـ	1	ذـ	1	0	0	0	0
/z/	زـ	1	زـ	1	0	0	0	0
/r/	رـ	1	رـ	1	0	0	0	0
/z/	زـ	1	زـ	0	0	0	0	0
/ʒ/	ژـ	0	ژـ	0	0	0	0	0
/s/	سـ	1	سـ	1	سـ	0	سـ	1
/ʃ/	شـ	1	شـ	1	شـ	0	شـ	1
/s/	صـ	1	صـ	1	صـ	0	صـ	1
/z/	ضـ	1	ضـ	0	ضـ	0	ضـ	0
/t/	طـ	1	طـ	1	طـ	0	طـ	0
/z/	ظـ	1	ظـ	0	ظـ	0	ظـ	0
/ʔ/	عـ	1	عـ	1	عـ	1	عـ	1
/q/	قـ	1	قـ	0	قـ	0	قـ	0
/f/	فـ	1	فـ	1	فـ	0	فـ	1
/G/	غـ	1	غـ	1	غـ	0	غـ	1
/k/	کـ	1	کـ	1	کـ	0	کـ	1
/g/	گـ	1	گـ	0	گـ	0	گـ	0
/l/	لـ	1	لـ	1	لـ	1	لـ	1
/m/	مـ	1	مـ	1	مـ	0	مـ	1
/n/	نـ	1	نـ	1	نـ	1	نـ	1
/v/	وـ	1	وـ	1	0	0	0	0
/h/	هـ	1	هـ	1	هـ	0	هـ	1
/j/	یـ	1	یـ	1	یـ	1	یـ	1

This experiment based on frequency is a sort of reverse engineering that helps us understand how Persian graphemes are encoded in a multilingual model like Whisper. Beyond this architectural bias based on the distributions of the subtokens in the Whisper multilingual model, we explore how the linguistic features of Persian are encoded in Whisper when transcribing actual Persian speech.

5.3. Experiment on Spontaneous Data

To better understand the phonographemic representation of Whisper, we analysed [the decoder output] of a 69-minute spontaneous video from YouTube. The context is a studio where three young men (in their 30s) discuss various social and cultural topics. The session is mostly informal, with frequent laughter, repetition, and other linguistic and non-linguistic discourse markers, which could pose challenges for ASR systems to transcribe.

To assess the accuracy of Whisper in transcribing this audio file, we report the Word Error Rate (WER) and Character Error Rate (CER). The WER of 0.403 suggests that approximately 40% of the words in the Whisper’s prediction are incorrect compared to the gold standard, indicating a significant discrepancy between the prediction and the reference. This is generally considered poor performance for many tasks, as nearly half of the words are incorrect. A CER of 0.222 suggests that 22.27% of the characters in Whisper’s prediction are incorrect when compared to the gold standard, reflecting a relatively high error rate in Whisper’s transcription.

TABLE 4. Performance Metrics for Prediction Accuracy of Whisper

<i>Metric</i>	<i>Value</i>
<i>Precision</i>	0.0004
<i>Recall</i>	0.0003
<i>F1-Score</i>	0.0003

The results for three other metrics Precision, Recall and F1-Score (see Table 4) are extremely low, indicating poor alignment between the Whisper’s predictions and the gold standard at the word level. Here is what each of these metrics suggests:

Precision measures how many of the predicted words are actually correct. A precision of 0.0354% means that out of all the words Whisper predicted, almost none of them were correct. This indicates a high number of false positives (words predicted incorrectly). Recall measures how many of the actual correct words were successfully predicted. A recall of

0.029% means the system is missing nearly all the correct words, suggesting a very high number of false negatives. F1-Score combines both precision and recall into a single metric. With such a low score, it reflects that both the precision and recall are very low, indicating a significant mismatch between the prediction and the reference.

5.4. The Speech to Subtoken Mapping

The analysis of the Persian subtokens derived from the dataset shows how in the distribution of token frequencies, certain characters and sequences appear much more frequently than others, as illustrated in Figure 5. The most frequent subtokens are key characters that are integral parts of many common Persian words. The dominance of these components reflects both the linguistic characteristics of Persian and the tokenisation strategy employed by the Whisper model.

Inventory of the most frequent Persian subtokens used in Whisper Predictions

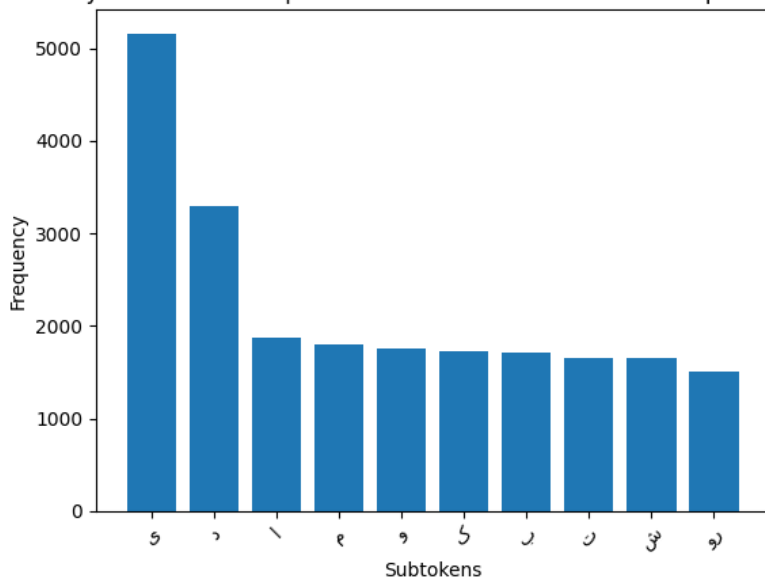


FIG. 5. Distribution of Persian Graphemes in our YouTube Dataset

Another important aspect of subtokenisation is the length of the subtokens across different languages. In the case of Persian, the tokenisation process tends to generate shorter subtokens. Based on our analysis of the subtokenisation of Persian YouTube data, out of a total of

30,544 subtokens generated by Whisper, 24,190 consist of only a single Persian character. Subtokens containing two Persian characters account for 4,325, while those with three Persian characters are limited to just 11. This observation suggests that subtokenisation for Persian, as an under-resourced language, may differ significantly from languages with larger training datasets, potentially reflecting the challenges and limitations in processing languages with less representation in model training.

5.5. Experiment on Whisper’s Alternate Predictions

We ran another experiment based on the experimentally customised implementation of C++ Whisper that extracts the 10 first predictions of the model and the corresponding probabilities¹². Do we get subtokens reserved for Persian or character-based representations of corresponding sounds for the alternate predictions?

	H		I		J		K		L		M		N		O		P		Q		R		S		T		U		V		W		X		Y		
1	Total tokens	Subtoken1	Prob 1	Subtoken2	Prob 2	Subtoken3	Prob 3	Subtoken4	Prob 4	Subtoken5	Prob 5	Subtoken6	Prob 6	Subtoken7	Prob 7	Subtoken8	Prob 8	Subtoken9	Prob 9	Subtoken10	Prob 10	Subtoken11	Prob 11	Subtoken12	Prob 12	Subtoken13	Prob 13	Subtoken14	Prob 14	Subtoken15	Prob 15	Subtoken16	Prob 16	Subtoken17	Prob 17	Subtoken18	Prob 18
1	67	[BEG_]	0.96207	L_TT_191	0.0025574	L_TT_191	0.0022283	L_TT_191	0.0020543	L_TT_191	0.00201671	L_TT_41	0.00174702	L_TT_141	0.00127791	L_TT_191	0.0017573	L_TT_191	0.0017573	L_TT_191	0.0017573	L_TT_191	0.0017573	L_TT_191	0.0017573	L_TT_191	0.0017573	L_TT_191	0.0017573	L_TT_191	0.0017573	L_TT_191	0.0017573	L_TT_191	0.0017573	L_TT_191	0.0017573
2	67	0.798019	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316	0.000316
3	67	0.908172	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	0.000912	
4	67	0.960564	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	0.0013278	
5	67	0.988139	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873	0.0006873		
6	67	0.669722	0.16727	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393	0.002393		
7	67	0.91857	0.054489	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867	0.017867		
8	67	0.94981	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379	0.002379		
9	67	0.99131	0.00234555	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495	0.00139495		
10	67	0.93413	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
11	67	0.941327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
12	67	0.9341327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
13	67	0.9341327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
14	67	0.9341327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
15	67	0.9341327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
16	67	0.9341327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
17	67	0.9341327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
18	67	0.9341327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
19	67	0.9341327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
20	67	0.9341327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
21	67	0.9341327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
22	67	0.9341327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
23	67	0.9341327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
24	67	0.9341327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
25	67	0.9341327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476		
26	67	0.9341327	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476	0.002476																						

FIG. 6. Top 10 alternate predictions for the sentence 001-A.wav file from the Common Voice corpus (large model)

Figure 6 shows the prediction used in the final transcription (prediction 1) and the alternate predictions; this is similar to the beam search for machine translation, providing access to the alternative transcriptions of the model. For instance, the first line (which predicts the special token), lists the alternate candidates for the transcription, all of which correspond to timestamps close to the initial starting point of the file. In the highlighted line of predictions (1.16), the first prediction (with a high level of certainty, 0.9969) corresponds to the unique letter that

12. <https://github.com/ggerganov/whisper.cpp>

corresponds to the bilabial nasal <م>. Predictions 2 and 4 correspond to combinations with the vowel <ما> and the consonant <مر>, respectively, while predictions 3 and 5 correspond to the Latin letter encoding the bilabial nasal. An unknown character appears in prediction 6. The seventh and eighth generated subtokens are <ب>, a voiced bilabial plosive, and <ن>, a voiced alveolar nasal. These cases are particularly intriguing as they highlight challenges faced by audio LLMs like Whisper, in distinguishing between phonemes with similar places or manners of articulation. This comparison of the Whisper transcription (prediction 1) with the alternate predictions allows us to investigate the subtokens that are the potential graphemic candidates that correspond to a portion of the acoustic signal. This technique allows us to probe the subtokens that may correspond to the corresponding speech interval. This method allows us to answer this phonographemic question : What are the competing subtokens for a given representation of the signal for a given time slice of the signal and the competing languages or the competing subtokens for a given portion of the signal?

6. Results

This section presents the key findings, including the qualitative and quantitative analyses from the transcription task using various Whisper models (subsections 6.1 and 6.2). Additionally, subsection 6.3 highlights the findings on subtokenisation phase in Whisper for Persian.

6.1. Qualitative Analysis of the Transcriptions

This subsection evaluates the performance of various Whisper models in transcribing Persian. The base model showcases robust multilingual outputs, such as tokens observed from other languages like Serbian, Chinese, Arabic, and even mixed-language phrases, as seen in *Après* (=after) *getting* (a blend of French and English). Furthermore, it often produces hallucinations in the form of meaningless, repetitive phrases—particularly in the middle of transcriptions—which affect the coherence and reliability of the output (Guerreiro et al., 2023).

The tiny model exhibits a higher frequency of hallucinations than the base model, often repeating clusters of characters or sentences. These hallucinations frequently contain words from other languages, predominantly English. Hallucinations are found in multilingual models in various translation contexts. Hallucinations are most common in low-resource languages and non-English output, sometimes exposing toxic patterns associated with training data (ibid.). Toxic patterns refer to

harmful or undesirable outputs exhibited by AI systems due to problems in their training data or algorithms.

The small model reduces the presence of foreign tokens significantly, with only isolated English words occasionally appearing within Persian transcriptions. This improvement indicates better language isolation, allowing for clearer and more coherent transcriptions with fewer instances of language interference.

In the medium model, Persian transcription accuracy improves further; however, the model fails to handle the same phoneme, i.e., /z/, which has multiple graphemic representations in Persian: <ز>, <ذ>, <ض>, and <ظ>. Additionally, there are instances where the sequence of Persian characters is rendered incorrectly, resulting in absurd or invalid forms. For example, the noun <لطف> (/lotf/, meaning *grace*) is transcribed incorrectly as <لوتف> (/lutf/), which is not a valid sequence in Persian. There are two issues with this transcription. First, the short vowel /o/, which is not explicitly written in Persian, is incorrectly transcribed as the long vowel /u/ in the erroneous version of the word. Second, the transcription selects the wrong grapheme to represent the phoneme /t/ in Persian, choosing <ت> instead of <ط>. Overall, while the model achieves greater accuracy in transcribing Persian, issues with hallucinations persist.

Finally, the large models offer the highest level of transcription quality, virtually eliminating hallucinations and foreign language tokens. Transcriptions are consistently accurate and free from unrelated phrases. Named Entity Recognition (NER) remains an issue. For example, the proper noun <رضا> (Reza) is written with the grapheme <ض>, which represents the /z/ sound. However, even though we would expect such a frequent first name to be present in the training data, the model consistently uses the grapheme <ز>, which is not used for the spelling of this proper noun.

Another significant factor influencing transcription, whatever model is used, is the presence of loanwords from languages such as Arabic and Urdu. These borrowed words, which are integrated into Persian vocabulary, may exhibit unique phonetic and morphological patterns that affect how the model processes and transcribes them. The handling of loanwords can highlight the model’s adaptability to cross-linguistic influences, but it also raises questions about the comprehensiveness of its training data in capturing these nuances. For example, <شکر> (/ʃokr/) means “gratitude” in both Arabic and Persian.

Moreover, the language frequency distribution within the Whisper dictionary plays a crucial role in shaping transcription outputs. The frequency of subtokens directly impacts the likelihood of their selection during the transcription process. If a subtoken appears more frequently in the training data, the model is more inclined to produce it, even in

contexts where a less frequent variant might be more appropriate. This distribution can influence both the selection of formal versus colloquial variants and the representation of loanwords, underscoring the interplay between token frequency and linguistic accuracy in Whisper’s transcriptions.

Last, a noteworthy observation across Whisper models is the addition of a letter referred to as the silent /he/ to represent the *ezafe* construction, which corresponds to an unstressed /e/ sound. In standard Persian orthography, this [e] sound is not explicitly transcribed. For example, Whisper transcribes <دوسته من> (/dusteh man/, meaning “my friend”), whereas the gold standard transcription in Persian is <دوست من> (/duste man/, meaning “my friend”). This discrepancy reflects a common challenge encountered by Persian learners and individuals with limited formal education, highlighting the difficulty of aligning phonetic and orthographic conventions in Persian transcription.

6.2. Quantitative Analysis of the Transcriptions

This subsection presents a statistical analysis of the transcriptions generated by Whisper, focusing on Word Error Rate (WER) and Character Error Rate (CER) as key metrics to evaluate transcription accuracy (Morris, Maier, and Green, 2004). For Persian, CER is particularly valuable, as it accounts for variations in spacing, which are common in the language and can impact transcription evaluation. Unlike WER, CER effectively ignores inconsistencies in how spaces are placed between letters, providing a more precise measure of accuracy for Persian text. For Persian, CER is very close to a subtoken error rate.

TABLE 5. WER and CER results for different models. The best-performing model (Large) is shown in bold characters

<i>Model</i>	<i>WER</i>	<i>CER</i>
Base	0.97	0.84
Tiny	2.66	1.22
Small	0.89	0.58
Medium	0.83	0.56
Large	0.42	0.20
Large-v2	0.57	0.24

The large Whisper model performed best, with the lowest WER (0.42) and CER (0.20). It captures both word-level and character-level accuracy effectively. The medium model also performed well, with WER (0.83) and CER (0.56), though it is not as accurate as the large model.

Base and small models showed higher WER and CER compared to both medium and large models. Tiny model performed the worst, with extremely high WER (2.66) and CER (1.22), indicating significant transcription errors. Model large-v2 strikes a balance but does not outperform the large model.

6.3. Distribution of Subtokens in Transcription Across Models

Our comparative analysis between model variants reveals hierarchical performance patterns aligned with model size. The larger models, particularly large and large-v2, demonstrate marginally superior vocabulary acquisition capabilities, especially beyond the 10^3 word threshold. The medium and small variants maintain remarkably similar growth trajectories through the middle ranges, while the tiny and base models show distinct divergence patterns at various intervals. This divergence becomes more pronounced in the upper word count ranges, suggesting that model capacity plays an increasingly significant role in vocabulary acquisition as the lexical complexity increases. Notably, several models exhibit plateau phases at different points in their growth curves, indicating potential capacity limitations or optimization boundaries specific to each model architecture.

In this subsection, we report our findings on the transcription of the data. Figure 7 displays the frequency of different Persian subtokens produced by the Whisper tiny, medium, and large-v1 models. The x-axis represents different Persian subtokens. The y-axis shows the frequency values, ranging up to 5000. As for the tiny model, the subtoken on the far left <ی> has the highest frequency, exceeding 5000. The remaining subtokens have lower frequencies, ranging approximately between 1000 and 2000. The pattern shows an interesting distribution, with one dominant token and several others at lower frequencies. The bar chart for the Medium Whisper model’s Persian subtoken distribution shows that the highest frequency is 4100 occurrences for the subtoken <ی>. The second most frequent subtoken <ه> has approximately 3750 occurrences. There is a significant drop after the second token, with the next group clustering around 1500-2000 occurrences. The lowest frequencies, <م> and <و>, are seen in the last two subtokens, with around 300-400 occurrences. In sum, the graph illustrates a clear non-uniform distribution of Persian subtokens. This distribution may affect how well the model handles different types of Persian text.

Figure 7 also shows the frequency of different Persian subtokens produced by the large-v1 model of Whisper. There appears to be a descending pattern in frequency from left to right like the other models. The most frequent subtoken <ی> has approximately 5000 occurrences,

and the second most common subtoken <د> has around 3300 occurrences. The remaining subtokens show relatively consistent frequencies between 1500-2000 occurrences. The distribution pattern is not uniform, showing a sharp decline between the two first frequent subtokens. After the second subtoken, the frequency stabilises with a more gradual decline.

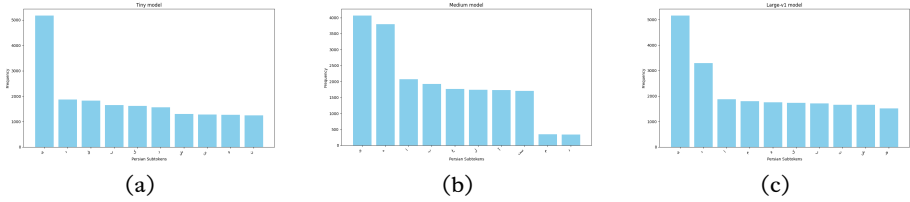


FIG. 7. Comparison of the distribution of subtokens across models

By analysing the distribution of subtokens in Whisper outputs across different Whisper models, we can compare these distributions to those of Persian subtokens within the FLEURS dataset, as illustrated in Figure 2. All Whisper models, with the exception of smaller models (e.g., tiny), demonstrate consistency with the distribution of graphemes in the FLEURS dataset. The graphemes <د> and <ا> are the most frequent in the FLEURS dataset, a pattern also reflected in the outputs of the larger Whisper models, namely medium and large-v1.

6.4. Probing of the Ten Best Predictions

In this experiment, we analysed the ten best predictions for the large model in 15 minutes of speech from the Mozilla Foundation. First, we show the validity of the results of our method, where we probe the alternative predictions for the special subtoken used to indicate the beginning of transcriptions [_BEG_]. We used the density probability representation (Figure 8), which is why the probabilities of the unique occurrences of [_TT_30] [_TT_6] are not represented. The most frequent second alternate predictions are [_TT_1] (5 occurrences) and [_TT_2] (23), where these special subtokens correspond to time stamps (Ballier, Burin, et al., 2024). Even if the time alignment is far from perfect with Whisper (ibid.), the results are consistent with the corresponding portion of the signal: these time stamps point to the very beginning of the signal (the bigger the [_TT_] value, the further from the beginning).

We similarly illustrate the consistency of the Whisper alternate predictions for the voiced alveolar fricative and for the bilabial nasal as case

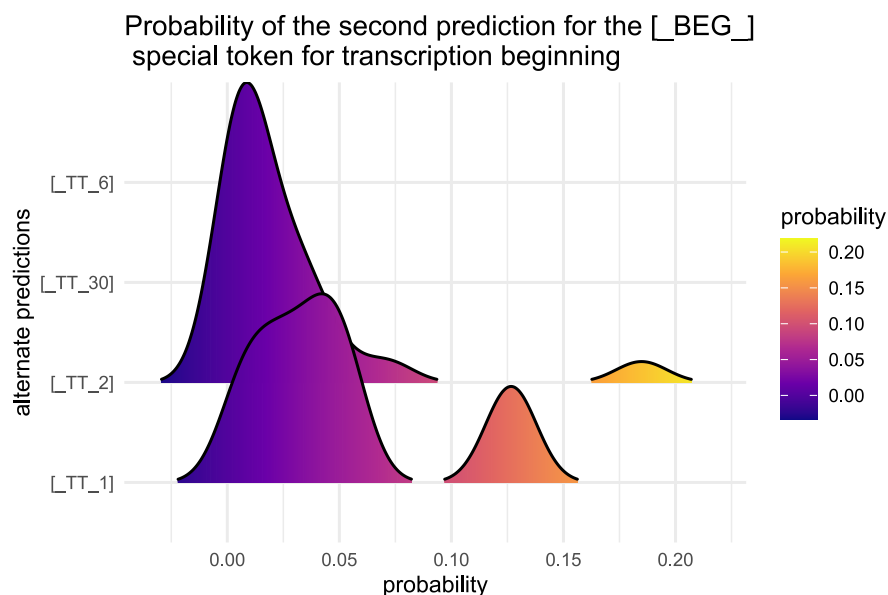


FIG. 8. The second best predictions for the [_BEG_] Whisper special token for transcription beginning (large model)

studies. For the transcription of the voiced fricative /z/ for the word <همزمان>, the expected letter would be <ز>. However, alternate predictions include subtokens belonging to other languages, like *zam* /zam/, which is a Russian word possibly meaning “deputy” or “assistant” or the subtoken in Latin characters *zaman* (see Figure 6). The first prediction corresponds to the correct letter in Persian for the voiced alveolar fricative. Nevertheless, the alternative letters are used in the predictions three and five and six and eight, but they are mostly used in Arabic. Alternatively, the full subtoken *zaman* is used in the fifth prediction and corresponds to the full realization of the following sounds. The alternate candidates for the same portion of the speech signal are nevertheless consistent rendering and transcriptions of the sibilant, exposing the competition of the subtokens of the different languages in the prediction. This probing of the alternate predictions confirms the correspondence between the token predicted and the acoustic properties of the corresponding speech signal.

A similar phenomenon can be observed with nasals. They all correspond to subtokens in different languages that correspond to the same phonological sounds like <m> and <M> with the Latin script. Figure 9 illustrates the alternate predictions for the bilabial nasal *min*, where the following vowel can be observed as well as the alveolar nasal <ن>.

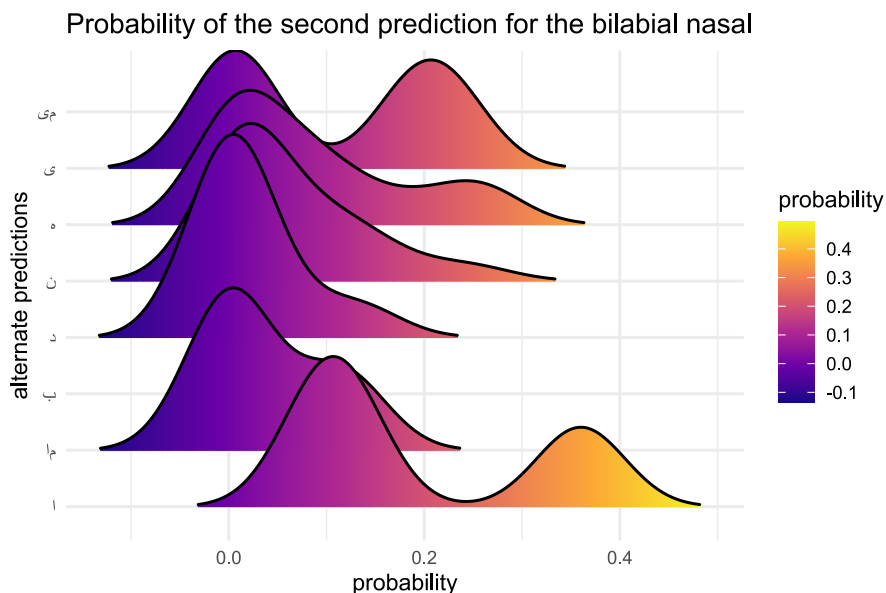


FIG. 9. The second best predictions for the bilabial nasal (large model)

Probing Whisper at the prediction phase does not capture the graphotactics of Whisper, but rather is a vantage point for the characterisation of the phonographematics of an audio LLM. It also efficiently summarises the dictionary biases of the subtokenisation: languages with a limited number of subtokens in the Whisper dictionary are more likely to be transcribed with subtokens used to represent other languages. Since all the Whisper models were trained with the same subtoken dictionary, this probably explains the hallucinations observed with the lower models: the transcriptions use ‘foreign’ subtokens that are deemed to be more probable, because the association of the foreign subtoken with the acoustic signal is probably more present in the training data of the other languages.

7. Discussion

7.1. Distribution of Subtokens in Whisper Dictionary Across the Perso-Arabic Script

We assume the byte pair encoding model (the subtokeniser) was trained on the full dataset of the Whisper training data, where Arabic

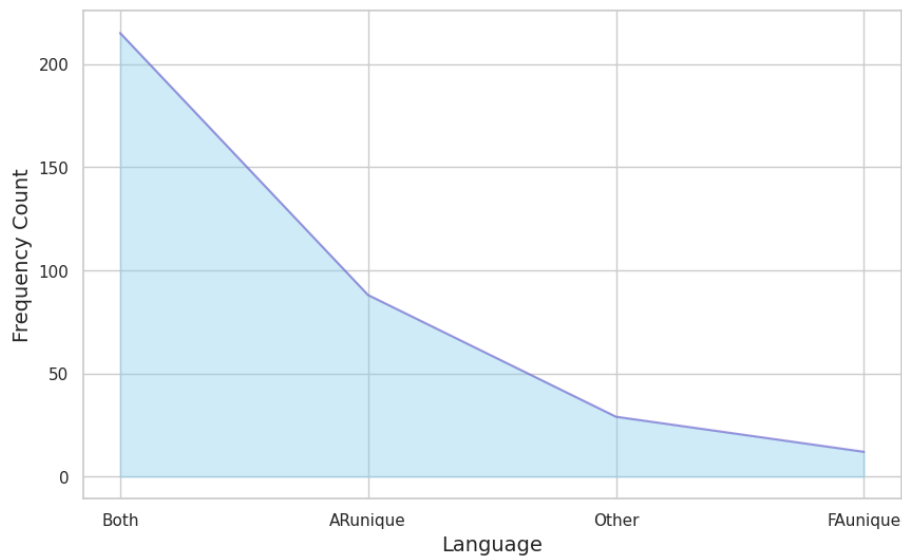


FIG. 10. Language Frequency Distribution in the Whisper Dictionary (large model)

has 739 hours and Persian 24 for the transcription task. To analyse the distribution of Perso-Arabic written characters in the Whisper multilingual dictionary, we extracted the corresponding subtokens [of the large model]. Figure 10 shows that 345 subtokens are associated with Perso-Arabic characters. Among these, 215 subtokens correspond to both Persian and Arabic (labeled as “Both”), 88 subtokens are specific to Arabic (labeled as “ARunique”), and only 12 subtokens are exclusive to Persian, representing characters that are uncommon in Arabic but unique to Persian (labeled as “FAunique”). Tadjiki and Urdu subtokens are labeled as “Other”.

Figure 10 highlights the over-representation of Arabic, which shares the same script as Persian, in Whisper’s subtoken dictionary. This imbalance explains the predominance of Arabic outputs, particularly in Whisper’s smaller models (e.g., base, tiny, and small). Other languages are also present in the dictionary, notably Chinese, which reflects the challenges of multilingual models with uneven language distributions. This issue has been studied, and prior research suggests that fine-tuning can significantly improve performance for under-represented languages by allocating more subtokens to them (Downey, Blevins, Goldfine, and Steinert-Threlkeld, 2023).

7.2. Sensitivity to Gender, Social Level, Accent, Register

Understanding how speaker attributes—such as gender, social background, accent, and register—affect ASR system performance is a critical area of investigation. Graham and Roll (2024a) and Patman and Chodroff (2024) focused on English, evaluates the impact of these variables on OpenAI’s Whisper ASR system. The results reveal Whisper’s strong recognition capabilities for North American English accents but highlight challenges with British, Australian, and certain non-native accents, particularly Vietnamese and Thai. Speaker-specific characteristics, including English proficiency, vowel inventory, and language typology, significantly influence recognition accuracy. Notably, male speakers generally experience higher error rates. Whisper also demonstrates superior performance with read speech compared to conversational speech, though the gender-related accuracy gap diminishes in conversational contexts. These findings underscore the importance of developing robust and inclusive ASR systems that can effectively handle diverse accents and speech patterns. Future research should prioritise reducing biases, utilizing more diverse datasets, and advancing personalized, context-aware ASR technologies to enhance real-world applicability. For Persian, building on previous experiments (Namdarzadeh et al., 2023), we report our findings on the variability of Whisper outputs for the same text read by two different native Tehrani speakers. Based on this experiment, Despite both female speakers employing the Tehrani dialect, commonly spoken in Iran’s capital, variations in the articulation of certain consonants and vowels were observed when the same word was read. Specifically, phonemic confusions were noted, including between /m/ and /b/, as well as /ʃ/ and /tʃ/. Additionally, inconsistencies arose in the differentiation between short and long vowels, such as /e/ (close-mid front unrounded) and /a/ (open front unrounded). These findings suggest that subtle differences in speakers’ manner of articulation can influence the transcription outputs, even for the large model.

7.3. The Subtokenisation Architectural Bias

An important factor to further investigate for the phono-graphematics of Whisper is the role of the distribution of languages and of the subtokenisation algorithm. Assuming the transcriptions of 680,000 hours of English are mixed all the other languages for the training of the byte pair encoding algorithm, under-resourced languages seem to be doomed to a character-based representation : whereas languages like English are represented by several thousand subtokens, some of them being actual words, less represented language only get a small fraction of the subtoken dictionary. The unequal distribution of the languages in the training

data lead to an architectural bias in the Whisper models (Ballier, Burin, et al., 2024). Previous research has pointed out some architectural biases of subtokenisation for NLP tasks. For example, for translation, Wisniewski, Zhu, Ballier, and Yvon (2021) have shown that the more subtokens are needed for occupational nouns, the more gender bias errors are likely to be found in translations. Therefore, it is likely that subtokenisation raises issues that are still to be uncovered, whether in the representation of the written language (Downey, Blevins, Goldfine, and Steinert-Threlkeld, 2023) or for the spoken language. Our paper is an initial investigation of these issues using a language that is under-represented in the lists of languages used for the training data, sharing letters from the same Unicode group, suggesting some of the dominant effects that can also be pointed out when investigating the transcriptions from Whisper. Clearly, the model size, therefore the number of representations, plays a crucial role, as well as, predictably enough, the size of the training data. Compared to the larger model where very few representations were spotted when transcribing 24 hours of speech of Persian, errors of that type could be spotted in the different models.

It may be the case that alternate (sub)tokenisation algorithms may be preferable, a point already made for the morphological tokenisation of Soranî Kurdish (Ahmadi, 2020), which is written in a modified Perso-Arabic script. Because of its specific orthographic and structural properties, the Perso-Arabic script presents specific challenges and opportunities for subtokenization. Subtokenisation models need to account for these differences to achieve effective morphological segmentation, and it would seem that the Unigram model tend to outperform BPE for in this task. It should be noted that this issue has been raised for LLMs (Petrov, La Malfa, Torr, and Bibi, 2024) and mitigating strategies (Downey, Blevins, Goldfine, and Steinert-Threlkeld, 2023; Han and Zou, 2024) have been proposed for written models, and comparable experiments should be carried out with audio LLMs.

8. Conclusion and Future Research

We have probed the graphemic outputs of Whisper and some of its internal representations for the transcriptions of spontaneous speech. Our reverse engineering perspective has enabled us to document how graphemes of Persian are encoded in a multilingual model like Whisper. Our methodology could be replicated for other languages that have a writing system potentially biased in the (sub)tokenisation phase. We have been able to discuss architectural bias and properties of subtokens and explore how linguistic properties (like the *ezafe*) are encoded in LLMs.

One of the questions is the generalisability of our reverse engineering method. We can suggest that it can be extended to all the languages which are treated by Whisper as a character-based model, namely languages for which the Whisper dictionary of subtokens mostly comprises their alphabet. As a result of which, the decoder has to apply most of the graphotactic constraints of the said language. This is what we showed with the automatic substitution of initial and final letters which were required in the final transcript as opposed to the grapheme in isolation that are predicted internally, see Table 3 that summarizes this point. The list of languages likely to be treated on a character-based system is likely to be deducted from the small number of hours in the training data as published in the (Radford et al., 2023) appendix.

The demographic of the speakers in the Whisper training data is only partially documented in Radford et al. (ibid.) since the training data is not really known, even for the small number of hours (24) of the Persian training data. As we suggested, we can nevertheless surmise that similar effects that have been diagnosed for English (Graham and Roll, 2024a; Sanabria et al., 2023) can be observed, for example, a different sensitivity to accents and gender has been documented for English (Graham and Roll, 2024a). Future research will need to see if the fine-tuned model for Persian using the Mozilla Foundation Common Voice data¹³ has similar gender bias for speech recognition.

Our strategy has been to treat the Whisper dictionary of subtokens as a repository of graphemic units mapped to acoustic representations of varying scope (from phonemic, syllabic to word-based representation for languages like English) correspondences.¹⁴ We have shown that Persian corresponds the character-based representation : a limited number of subtokens in the dictionary was used in the encoder for the prediction phase and graphotactic rules are applied by the decoder.

More generally, our roadmap for investigating the phonological “knowledge” of audio LLMs consists of investigating the existence of phonographemic mapping rules for subtokens. For this phonology of subtokens, if we assume that subtokens are associated to similar pronunciations, what is the acoustic granularity of this similarity? In the name of robustness, we have to assume some plasticity for the acoustic input matching the subtoken, a unique subtoken corresponding to several phonetic realisations. A systematic analysis of the top ten alternate predictions (see Figure 6) for the transcription task should provide insight into the phonetic representations of the different models: the subtokens

13. <https://huggingface.co/speechbrain/asr-whisper-large-v2-commonvoice-fa>

14. We do analyse the dictionary of the English only models, based on GPT2, nor do we detail here the fraction of special subtokens used to represent tasks, task position of time stamps, see for instance Ballier, Burin, et al. (2024).

predicted reflect the categorisation of phonetic neighbours as encoded in the Whisper subtoken representations.

9. Acknowledgements

The authors would like to express their gratitude to Prof. Richard Wright for his assistance with our revised phonological analysis of Persian and the phonetic correlates potentially associated with Persian graphemes. Additionally, they are grateful to Dr. Jean-Baptiste Yunès for providing the C++ implementation of OpenAI’s Whisper, which made this work possible. The authors would like to extend their appreciation to the reviewers for their insightful comments and constructive feedback, which have significantly enhanced the quality of this work. We also thank Guillaume Wisniewski and Alice Henderson for comments on an earlier version of this paper.

A. Appendix

TABLE 6. Perso-Arabic Subtokens in the Multilingual Whisper Dictionary

ك	س	ة	ع	ب	ه	د	ت	و	م	ر	ن	ي	ل	ا
ك	في	ذ	ن	ش	ه	ى	ج	ي	ع	ال	أ	ف	ح	ق
ز	ون	ح	ها	ض	عل	ام	ج	ية	ين	خ	الم	ار	ط	ص
على	ذا	ما	لا	ست	أن	وا	الل	ير	ور	ث	ر	ق	إ	د
كن	ما	عن	الع	اد	غ	گ	اب	الت	ري	خ	ان	دي	الأ	هم
ط	ني	عد	الله	الح	بي	الس	بال	وم	مع	لي	ئ	ظ	آ	لم
رب	الش	الق	لل	كم	اس	حد	ذه	چ	هذا	را	ود	لك	كى	
									علي	ال	رة	يد	يا	

Among the 344 Perso-Arabic characters extracted from the Whisper dictionary, 240 are Persian but not exclusive to Persian, as they are also used in Arabic. Punctuation marks are recorded in the Perso-Arabic format, following the right-to-left writing system.

The remaining characters belong to Arabic transcripts, some of which have been phased out of Persian due to the language’s evolution and domestication.

References

- Ahmadi, Sina (2020). "A tokenization system for the Kurdish language". In: *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pp. 114–127.
- Arab-Moghaddam, Narges and Monique Sénéchal (2001). "Orthographic and phonological processing skills in reading and spelling in Persian/English bilinguals". In: *International Journal of Behavioral Development* 25, pp. 140–147. DOI: 10.1080/01650250042000320.
- Ardila, R. et al. (2020). "Common Voice: A Massively-Multilingual Speech Corpus". In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 4211–4215.
- Asebriy, Zahra et al. (2014). "Comparative systems of handwriting Arabic character recognition". In: *2014 Second World Conference on Complex Systems (WCCS)*, pp. 90–93. DOI: 10.1109/ICoCS.2014.7060923.
- Ballier, Nicolas, Léa Burin, et al. (2024). "Probing Whisper Predictions for French, English and Persian Transcriptions". In: *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*. Ed. by Mourad Abbas and Abed Alhakim Freihat. Trento: Association for Computational Linguistics, pp. 129–138. URL: <https://aclanthology.org/2024.icnlsp-1.15>.
- Ballier, Nicolas et al. (2023). "Using Whisper LLM for Automatic Phonetic Diagnosis of L2 Speech, a Case Study with French Learners of English". In: *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*. Ed. by Mourad Abbas and Abed Alhakim Freihat. Online: Association for Computational Linguistics, pp. 282–292. URL: <https://aclanthology.org/2023.icnlsp-1.30>.
- Baluch, Bahman and S. Choudhury (2011). "Spelling transparency and its impact on short-term memory: Evidence from persian and english". In: *Spelling Skills: Acquisition, Abilities, and Reading Connection*, pp. 93–106.
- Bostrom, Kaj and Greg Durrett (2020). "Byte Pair Encoding is Suboptimal for Language Model Pretraining". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4617–4624.
- Conneau, Alexis et al. (2022). "FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech". preprint arXiv:2205.12446.
- (2023). "Fleurs: Few-shot learning evaluation of universal representations of speech". In: *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, pp. 798–805.
- Devlin, Jacob et al. (2019). "BERT: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186.
- Downey, C.m. et al. (2023). "Embedding Structure Matters: Comparing Methods to Adapt Multilingual Vocabularies to New Languages".

- In: *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*. Ed. by Duygu Ataman. Singapore: Association for Computational Linguistics, pp. 268–281. DOI: 10.18653/v1/2023.mrl-1.20.
- Freihat, Abed Alhakim and Mourad Abbas, eds. (2021). *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Trento, Italy: Association for Computational Linguistics. URL: <https://aclanthology.org/2021.nsurl-1.0>.
- Gerganov, Georgi (2003). “whisper.cpp : A High-performance inference of OpenAI’s Whisper automatic speech recognition (ASR) model”. URL: <https://github.com/ggerganov/whisper.cpp>.
- Graham, Calbert and Nathan Roll (2024a). “Evaluating OpenAI’s Whisper ASR: Performance analysis across diverse accents and speaker traits”. In: *JASA Express Letters* 4.2, p. 025206.
- (2024b). “Evaluating OpenAI’s Whisper ASR: Performance analysis across diverse accents and speaker traits”. In: *JASA Express Letters* 4.2, p. 025206. ISSN: 2691-1191. DOI: 10.1121/10.0024876.
- Guerreiro, Nuno M. et al. (2023). “Hallucinations in Large Multilingual Translation Models”. In: *Transactions of the Association for Computational Linguistics* 11, pp. 1500–1517. DOI: 10.1162/tacl_a_00615.
- Halbout, Dominique and Muhammad-hossein Karimi (2012). *Le persan*. Assimil.
- Han, Yujin and Difan Zou (2024). “Understanding and Mitigating Tokenization Bias in Language Models”. In: *Proceedings of the 41st International Conference on Machine Learning*. Ed. by Ruslan Salakhutdinov et al. Vol. 235. Proceedings of Machine Learning Research. PMLR, pp. 17480–17504. URL: <https://proceedings.mlr.press/v235/han24g.html>.
- Hosseini, Fatemeh Sadat et al. (2021). “IDPL-PFOD: An Image Dataset of Printed Farsi Text for OCR Research”. In: *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Ed. by Abed Alhakim Freihat and Mourad Abbas. Trento, Italy: Association for Computational Linguistics, pp. 22–31. URL: <https://aclanthology.org/2021.nsurl-1.4>.
- Jafari, Mohammad Mahdi et al. (2021). “Improving pre-trained Language Model for Relation Extraction Using Syntactic Information in Persian”. In: *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Ed. by Abed Alhakim Freihat and Mourad Abbas. Trento, Italy: Association for Computational Linguistics, pp. 38–44. URL: <https://aclanthology.org/2021.nsurl-1.6>.
- Karimi, Sarvnaz, Falk Scholer, and Andrew Turpin (2007). “Collapsed Consonant and Vowel Models: New Approaches for English-Persian Transliteration and Back-Transliteration”. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Ed. by Annie Zae-

- nen and Antal van den Bosch. Prague, Czech Republic: Association for Computational Linguistics, pp. 648–655. URL: <https://aclanthology.org/P07-1082>.
- Karimi, Simin (2005). *A Minimalist Approach to Scrambling. Evidence from Persian*. Berlin, New York: De Gruyter Mouton. DOI: doi : 10 . 1515 / 9783110199796.
- Khalilia, Hadi, Abed Alhakim Freihat, and Fausto Giunchiglia (2021). “The Dimensions of Lexical Semantic Resource Quality”. In: *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Ed. by Abed Alhakim Freihat and Mourad Abbas. Trento, Italy: Association for Computational Linguistics, pp. 15–21. URL: <https://aclanthology.org/2021.nsurl-1.3>.
- Kudo, T. (2018). “Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing”. preprint arXiv:1808.06226.
- Kummervold, Per E et al. (2024). “Whispering in Norwegian: Navigating Orthographic and Dialectic Challenges”. preprint arXiv:2402.01917.
- Mahootian, Shahrzad and Lewis Gebhardt (1997). *Persian*. Descriptive grammars. Includes index. London: Routledge.
- Maleki, Jalal and Lars Ahrenberg (2008a). “Converting Romanized Persian to the Arabic Writing Systems”. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*. Ed. by Nicoletta Calzolari et al. Marrakech, Morocco: European Language Resources Association (ELRA). URL: <https://aclanthology.org/L08-1315/>.
- (2008b). “Converting Romanized Persian to the Arabic Writing Systems”. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*. Marrakech, Morocco: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/738_paper.pdf.
- Manova, Stela (2023). “ChatGPT, *n*-grams and the power of subword units: The future of research in morphology”. In: *Fourth International Workshop on Resources and Tools for Derivational Morphology*. Vol. 5, p. 1.
- Manova, Stela et al. (2020). “What is in a morpheme? Theoretical, experimental and computational approaches to the relation of meaning and form in morphology”. In: *Word Structure* 13.1, pp. 1–21.
- “Frontmatter” (2023). In: *Persian Computational Linguistics and NLP*. Ed. by Katarzyna Marszałek-Kowalewska. Berlin, Boston: De Gruyter Mouton, pp. I–IV. DOI: doi:10.1515/9783110619225-fm.
- Mohebbi, Hosein et al. (2023). “Homophone Disambiguation Reveals Patterns of Context Mixing in Speech Transformers”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Asso-

- ciation for Computational Linguistics, pp. 8249–8260. DOI: 10.18653/v1/2023.emnlp-main.513.
- Morris, Andrew Cameron, Viktoria Maier, and Phil D Green (2004). “From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition”. In: *Interspeech*, pp. 2765–2768.
- Namdarzadeh, Behnoosh et al. (2023). “Fine-tuning MBART-50 with French and Farsi data to improve the translation of Farsi dislocations into English and French”. In: *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*. Ed. by Masaru Yamada and Felix do Carmo. Macau SAR, China: Asia-Pacific Association for Machine Translation, pp. 152–161. URL: <https://aclanthology.org/2023.mtsummit-users.14/>.
- Oji, Romina, Nasrin Taghizadeh, and Heshaam Faili (2021). “PerSpell-Data: An Exhaustive Parallel Spell Dataset For Persian”. In: *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Ed. by Abed Alhakim Freihat and Mourad Abbas. Trento, Italy: Association for Computational Linguistics, pp. 8–14. URL: <https://aclanthology.org/2021.nsurl-1.2>.
- Patman, Chloe and Eleanor Chodroff (2024). “Speech recognition in adverse conditions by humans and machines”. In: *JASA Express Letters* 4.11.
- Petrov, Aleksandar et al. (2024). “Language model tokenizers introduce unfairness between languages”. In: *Advances in Neural Information Processing Systems* 36.
- Radford, Alec et al. (2023). “Robust Speech Recognition via Large-Scale Weak Supervision”. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 28492–28518. URL: <https://proceedings.mlr.press/v202/radford23a.html>.
- Rahbari, Noriyeh and Monique Sénéchal (2009). “Lexical and nonlexical processes in the skilled reading and spelling of Persian”. In: *Reading and Writing* 22.5, pp. 511–530. DOI: 10.1007/s11145-008-9122-1.
- Rastorgueva, Vera Sergeevna and Steven P. Hill (1964). *A Short sketch of the grammar of Persian*. Indiana University Research Center in Anthropology, Folklore and Linguistics Publication. “Also : Part II of the International Journal of Ammerican Linguistics”. Bloomington: Indiana University.
- Salehi, Ali and Cassandra L. Jacobs (2024). “The Effect of Model Capacity and Script Diversity on Subword Tokenization for Sorani Kurdish”. In: *Proceedings of the 21st SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*. Ed. by Garrett Nicolai et al. Mexico City, Mexico: Association for Computational Linguistics, pp. 51–56. DOI: 10.18653/v1/2024.sigmorphon-1.6.

- Sanabria, Ramon et al. (2023). "The Edinburgh international accents of English corpus: Towards the democratization of English ASR". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5.
- Sartakhti, Moein Salimi, Romina Etezadi, and Mehrnoush Shamsfard (2021). "Improving Persian Relation Extraction Models By Data Augmentation". In: *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Ed. by Abed Alhakim Freihat and Mourad Abbas. Trento, Italy: Association for Computational Linguistics, pp. 32–37. URL: <https://aclanthology.org/2021.nsurl-1.5>.
- Sennrich, Rico, Barry Haddow, and Alexandra Birch (2015). "Neural Machine Translation of Rare Words with Subword Units". In: *CoRR* abs/1508.07909. URL: <http://arxiv.org/abs/1508.07909>.
- (2016). "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, pp. 1715–1725. DOI: 10.18653/v1/P16-1162.
- Shakeri, N. et al. (2015). "Investigating Phonological Awareness in Persian-Speaking Children With Phonological Disorders". In: *Middle East Journal of Rehabilitation and Health* 2.4, e32200. DOI: 10.17795/mejrh-32200. URL: <https://brieflands.com/articles/mejrh-21516>.
- Song, Xinying et al. (2021). "Fast wordpiece tokenization". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2089–2103.
- Sproat, Richard and Alexander Gutkin (2021). "The taxonomy of writing systems: How to measure how logographic a system is". In: *Computational Linguistics* 47.3, pp. 477–528.
- Taghizadeh, Nasrin, Ali Ebrahimi, and Heshaam Faili (2021). "NSURL-2021 Shared Task 1: Semantic Relation Extraction in Persian". In: *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*. Ed. by Abed Alhakim Freihat and Mourad Abbas. Trento, Italy: Association for Computational Linguistics, pp. 1–7. URL: <https://aclanthology.org/2021.nsurl-1.1>.
- Taguchi, Chihiro and David Chiang (2024). "Language Complexity and Speech Recognition Accuracy: Orthographic Complexity Hurts, Phonological Complexity Doesn't". preprint arXiv:2406.09202.
- Wisniewski, Guillaume et al. (2021). "Screening Gender Transfer in Neural Machine Translation". In: *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 311–321.
- Yousef, Saeed and Hayedeh Torabi (2018). *Persian*. Routledge comprehensive grammars. Abingdon, Oxon: Routledge.

Developing Orthographic Conventions of Roman Hindi-Urdu

Saiya Karamali

Abstract. The Roman script has become ubiquitous over the past few decades for writing Hindi-Urdu online. Not only have the orthographic conventions of Roman Hindi-Urdu not been studied in detail, but the lack of prescriptive rules gives us an opportunity to examine how a writing system might develop organically. After presenting a brief history of Roman Hindi-Urdu, and discussing possible reasons for its popularity, I use data from X (formerly Twitter) to identify the most common orthographic representations for each Hindi-Urdu phoneme. I then investigate how consistent these conventions are and which psychological, sociocultural, and technological factors appear to have shaped their development. I find a remarkable tendency toward a one-to-one relationship between orthographic representations and phonemes, but I also find some considerations which appear to override this tendency. By documenting the orthographic conventions of Roman Hindi-Urdu, this paper provides a foundation for future work on Roman Hindi-Urdu as well as a possible example of what an organically developing writing system might look like.

1. Introduction

Written language is full of the same kinds of complexities and intrigue found in spoken language. In our modern era of frequent internet communication, written language is a medium for users of vastly different backgrounds to interact with each other, and users find ways of conveying social meaning through their choice of words, script, and orthographic conventions. The study of written language has been largely neglected in linguistics, and this paper is an attempt to fill that gap.

Advances in technology over the last two to three decades have accelerated the evolution of many writing systems throughout the world. This includes shorthand used for instant messaging, visual characters such as emojis, the development of orthographies for previously unwritten lan-

Saiya Karamali  0009-0007-8598-2769
University of Washington
E-mail: karamali@uw.edu

Y. Haralambous (Ed.), *Grapholinguistics in the 21st Century 2024. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2681-8566, e-ISSN: 2534-5192), Vol. 11.
Fluxus Editions, Brest, 2026, pp. 361-391. <https://doi.org/10.36824/2024-graf-kara>
ISBN: 978-2-487055-10-0, e-ISSN: 978-2-487055-11-7

guages, and the application of novel scripts to languages in which they were not previously used (Meletis, 2018; Meletis and Dürscheid, 2022).

Two languages that have been significantly affected by changing writing systems are Hindi and Urdu. These closely related Indo-Aryan languages are extremely similar in informal registers. In formal speech, lexical differences grow, but syntax and phonology remain very similar (Masica, 1991). Yet, Hindi and Urdu are generally considered separate languages for historical, political, and sociocultural reasons (Ahmad, 2011), and one of the most visible differentiators between Urdu and Hindi is their use of separate writing systems. However, in recent years, speakers of both languages have begun to use the Roman (Latin) script in online and mobile communication as well as on billboards and storefronts (Atta, 2021; Bali, Sharma, Choudhury, and Vyas, 2014). Since the languages are so similar and are often indistinguishable in this form, I consider them together, and will refer to them as Hindi-Urdu.

Although Roman Hindi-Urdu dates to British colonization of South Asia, it lacks prescribed spelling rules and, in the technological age, has undergone rapid adoption and evolution. In this paper, I use a dataset harvested from X (formerly Twitter) to determine how the various sounds of Hindi-Urdu are represented in Roman orthography, and to perform an analysis of what this tells us about Hindi-Urdu and about how writing systems tend to develop when given the chance to do so organically.

In Section 1, I summarize the background literature on writing systems, Roman Hindi-Urdu, the history of Hindi and Urdu, and Hindi-Urdu phonology. In Section 2, I describe the goals, dataset, and methods of this study. In Section 3, I examine the orthographic conventions for writing each phoneme in Hindi-Urdu. In Section 4, I present the entire list of correspondences between phonemes and graphemes and identify the considerations which appear to have shaped the orthographic conventions of Roman Hindi-Urdu. And in Section 5, I present several important avenues for future work.

1.1. Literature Review

1.1.1. *Terminology for Discussing Writing Systems*

Following Meletis and Dürscheid (2022), I consider a *grapheme* to be the smallest unit of a writing system. For an alphabet, that is equivalent to a letter. A *script* is the set of graphemes used for a language, and an *orthography* consists of a set of rules that govern how graphemes are put together to form words. A *writing system*, then, consists of a script and, optionally, an orthography.

1.1.2. *The History of Hindi and Urdu*

Several authors have posited that Hindi and Urdu originated from among many social dialects spoken in Delhi in the 1600s (Masica, 1991; Schmidt, 2003). The *Devanagari* script, derived from older scripts which developed on the Indian subcontinent (Masica, 1991), and Perso-Arabic script, likely brought to India by Persian invaders in the eleventh century, were both in use by then (Masica, 1991; T. Rahman, 2011; Salomon, 2003, pp. 58, 89). Starting in the 18th century, sociocultural divisions between Hindus and Muslims grew—often due to deliberate actions by British colonizers (Faruqi, 2003). By the twentieth century, the name ‘Urdu’ and the Perso-Arabic script had become associated with Muslims; and the name ‘Hindi’ and the Devanagari script had become associated with Hindus (T. Rahman, 2011, p. 99). This divide continued to grow after India and Pakistan became independent, with Pakistan adopting Urdu as its national language and India adopting Hindi.

The difference in writing systems remains the most salient across-the-board difference between Urdu and Hindi (Ahmad, 2011; Rao, Vaid, Srinivasan, and Chen, 2011). In spoken language, the two are mutually intelligible, and regional differences are often greater than the differences between Urdu and Hindi (Schmidt, 2003). But although both languages include Perso-Arabic and Sanskrit loans, Urdu can be characterized by the use of more Perso-Arabic loans, and Hindi by the use of more Sanskrit loans (Kachru, 2006; Shapiro, 2003). This is true of sounds, too, where Urdu speakers are more likely to have acquired the Perso-Arabic loans [q x y z f ʃ], and many Hindi speakers have acquired the Sanskrit loans [ɳ ʃ ʂ] (Schmidt, 2003). These differences are probably not entirely sociocultural; for instance, many Hindi speakers also may distinguish /f ʃ z/ (ibid.). Additionally, many self-identified Urdu speakers do not use the Perso-Arabic loans, or use them less frequently, and the same is true for self-identified Hindi speakers with Sanskrit loans (Fatima, 2011; Mangrio and Ali, 2014; Ohala, 1983). Still, these sounds are another possible way that speakers might distinguish Urdu from Hindi.

1.1.3. *History and uses of Roman Hindi-Urdu*

Roman Hindi-Urdu was used as early as the nineteenth century, particularly by European missionaries and British Indians (Ahmad, 2011; Nema and Chawla, 2018). The earliest Roman Hindi-Urdu publication with which I am familiar is Gilchrist’s (1803) multilingual translation of Aesop’s Fables, which includes Persian, Arabic, Hindi-Urdu, and several other South Asian languages, all written in the Roman script. Many Hindi-Urdu pedagogical materials since that time have used the Roman script as well, which may be because they were geared towards British learners or due to lower cost of publishing (e.g., Khan, 2000; A. Rah-

man, 1923; R. N. Sharma, 1937). Early Roman Hindi-Urdu publications generally devised their own systems of one-to-one correspondences between Roman characters and Hindi-Urdu sounds, and used a number of diacritics to do so. Some conventions from this period, such as the use of ⟨kh⟩ and ⟨gh⟩ to represent /x/ and /ɣ/, appear to have carried over to the present day, but many others—such as the representations for the vowels—have changed.

In general, Indians and Pakistanis under British rule did not adopt the Roman script (Faruqi, 2003), and upon independence in 1947, India and Pakistan officially adopted the traditional Hindi and Urdu scripts, respectively. Why, then, has Roman Hindi-Urdu exploded in popularity in recent years? Certainly, the emergence of technology has introduced the need to type, which was not always possible in scripts other than the Latin script. But with the introduction of physical and virtual keyboards for many more scripts, facility of use probably does not tell the whole story. Recent statistical analyses of Romanization are sparse, but languages appear to differ in whether the Roman script has continued to be used: for instance, Androutsopoulos (2009) and Mouresioti and Terkourafi (2021) both found that Romanization is unpopular among Greek speakers, but Bali, Sharma, Choudhury, and Vyas (2014) found that 84 percent of a sampling of Hindi Facebook posts were written in the Roman script.

A possible reason for the popularity of the Roman script in South Asia today is the prestige of English—which uses the Roman Script—in that region. English was introduced by the British, and was taught in elite schools dating back to the 19th century (T. Rahman, 2020). And despite rising opposition to English at the time of independence, it still became an official language of both India and Pakistan (National Assembly of Pakistan, 1973; Parliament of India, 1963) as well as a *lingua franca* in multilingual regions of both countries (Mukherjee, 2010; Zaidi and Zaki, 2017). English-language education remains widespread in private schools in both India and Pakistan, leading many in the middle and upper classes to be comfortable using both the English language and the Roman script in formal settings (T. Rahman, 2020; Ramanathan, 2016).

Starting in the 1990s, economic development and globalization have also contributed to increased prestige of English in South Asia, to the point where it has become “the’ language of Indian youth” (Nema and Chawla, 2018, emphasis in original). It is reasonable to suppose, then, that the script of English has gained prestige along with the language. Previous linguistic work has found concrete uses for Roman Hindi-Urdu, as well. Rosowsky (2010) observed a Pakistani immigrant community in the UK and found that the Roman script is often used for teaching Urdu and for teaching the Quran, and that the youth often use the Roman script to transcribe Urdu songs and poetry. For these youth, the Roman script is a tool to help preserve cultural, linguistic, and re-

ligious identity. Atta (2021) found that Roman Urdu is commonly used in billboards and storefronts in Pakistani cities, citing citizens' 'hybrid identity' due to colonization and globalization.

Thus, today's Roman Hindi-Urdu appears to be neither a colonial imposition nor a direct reaction to technology. Instead, it is a response to the modern needs and attitudes of many Hindi-Urdu speakers, and appears to be an increasingly important part of their linguistic repertoire.

1.1.4. *Writing Systems*

Previous studies have examined how people adapt new scripts to their own languages. In some cases, users have clearly articulated philosophies on how they do this: Androutsopoulos (2009) found that although Greek has no prescriptive standards for Romanization, most users used one of three loose 'systems': correspondence between Roman letters and Greek sounds; correspondence between the shapes of Roman and Greek letters; or correspondence between English and Greek keyboard locations. The second and third of these would then restrict possible readers to those with knowledge of Greek script. In other cases, spelling conventions are much less standardized: Akbari (2013) studied Romanized Persian text messages by Iranian users and found a wide variety of spelling techniques used, including multiple graphemes being used to represent the same phoneme, variation in whether to use digits to represent a sequence of characters, and variation in whether to omit vowels.

Sociocultural factors have also led to the use of new scripts and orthographies. Sebba (2007) describes cases of oppressed groups developing their own writing conventions to distinguish their writing from that of an oppressing group, as well as of graffiti artists using intentional "misspellings" to invoke humor or cause their work to stand out. And QDJY is an alternative orthography for Bahasa Indonesia which was deliberately created by an online community (Putra and Triyono, 2018). QDJY uses the same script as the language's official writing system, but its highly innovative spelling conventions serve to identify QDJY users as part of that community.

Meletis (2018) identifies four factors which may affect which script and orthographic conventions eventually become widespread for a particular language. The first of these is *linguistic fit*, or how well a writing system reflects the language. Generally, this is determined by how close to one-to-one the correspondence between sounds and their representations is. The second factor is *psychological* or *cognitive fit*, or the relationship between a writing system and our psychological, physical, and cognitive faculties. Previous studies such as Rao, Vaid, Srinivasan, and Chen (2011) have shown that some writing systems are more easily processed cognitively than others. Higher linguistic fit will often lead to higher psychological fit, but psychological fit can be measured independently

through empirical experiments of reading and writing ability. The third factor is *sociocultural fit*, or what attitudes users have about a script and how well it fits their identities. The fourth possible factor is *technological fit*, or how easily the script can be used on computers and mobile devices. Meletis posits that these factors shape how likely a script is to be used for a particular language.

The Devanagari script has the highest linguistic fit of the writing systems used for Hindi-Urdu. That script has close to a one-to-one correspondence between sounds and their representations, whereas the Urdu script has sounds which are represented by multiple symbols, symbols which represent multiple sounds, and at least one silent letter; and vowels are generally omitted. The Hindi writing system has a higher psychological and cognitive fit, as well: Rao, Vaid, Srinivasan, and Chen (2011) found that Hindi-Urdu biliterate readers were consistently able to read more quickly in Devanagari than in Perso-Arabic.

With regard to sociocultural fit, Unseth (2005) identifies three sociocultural aims that script choice can achieve: identification with a particular group; distancing from a particular group; and ease of interaction with outside communities. Indeed, the Perso-Arabic script has historically been so closely identified with South Asian Muslims that it has made them resistant to using both the Devanagari and Roman scripts, to the point where Urdu speakers sometimes use Perso-Arabic writing conventions even when writing in the Devanagari script, where such conventions are of no linguistic use (Ahmad, 2011). Additionally, the Roman script was historically associated with colonization, from which South Asians may have wished to distance themselves (*ibid.*). Recently, however, the Roman script has become associated with economic prosperity in South Asia (Nema and Chawla, 2018), which may have increased the sociocultural fit of the Roman script in both Hindi and Urdu. And finally, the Roman script may have the highest technological fit, because it was among the first scripts used on computers and mobile devices; but perhaps not overwhelmingly so, given that Urdu and Hindi keyboards are also available, particularly on mobile devices.

2. Methods

2.1. Dataset

I obtained my dataset by taking a random sampling of posts using the X streaming API¹, which returns a random sample of posts which are created while the stream is running. I collected posts between May 1

1. <https://developer.x.com/en/docs/tutorials/stream-tweets-in-real-time>

and May 9, 2023, resulting in a total of 56.5 GB of posts being collected. I filtered the posts to include only those in ASCII characters which were classified by X as either Hindi or Urdu. Web addresses and mentions were removed from posts, and duplicate reposts and quote posts were removed from the dataset. The resulting dataset consists of 8909 posts, only a small fraction of the posts collected.

More posts could not be collected because X removed access to its APIs for academic research as of June 8, 2023. X and other online fora are key sources of emerging linguistic data and unfortunately, none of these are currently available for reasonable pricing. X, for instance, now charges \$5000 per month to perform the data collection carried out in this study. However, the 8909 posts provide a sample large enough to perform a linguistic analysis of what spelling conventions are used, and which I supplement with more fine-grained qualitative analysis.

2.2. Methods

From the 8909 posts I retained from X, I extracted the 2000 most frequent orthographic “words”, where a word is a group of characters separated by whitespace. No effort was made at this stage to group together alternate orthographic representations which correspond to the same semantic word. I did, however, remove words which I judged to be proper names or loanwords from English, since proper names may have more standardized spellings and the spellings of English loanwords could be influenced by their English orthography.

In order to analyze each phoneme, I first used my own knowledge as well as manual examination of posts to identify the possible orthographic representations of each phoneme. I then filtered the set of orthographic words to include only words with these representations, and then manually inspected the resulting data and removed words which appeared not to include the target phoneme. I did this using my own knowledge as an Urdu speaker to determine the most likely phonetic form for each word, with assistance from Wiktionary (2023), where needed. Wiktionary has been previously used as a source of community-built transcriptions for a given community’s language variety (see Meyer and Gurevych, 2012; Noruzi, 2023), and appeared to be the most reliable source available. I then counted the occurrences of each orthographic representation for the given phoneme.

A few phonemes were too rare to exhibit meaningful patterns within the top 2000 words. In these cases, as noted in Section 3, I expanded the dataset by searching directly for possible representations of words which contain the target phoneme. I identified these words by using wordfreq (Speer, 2022) to list the most frequent Hindi words containing a given Devanagari character.

In Section 3, I present my findings for each phoneme in the Hindi-Urdu phonemic inventory. I discuss patterns in how various natural classes of phonemes are written and explore what these tell us about Roman Hindi-Urdu writing conventions. I then explore the possible effects of phonological variation, sociocultural variation, and perception of the various sounds by Hindi-Urdu speakers.

3. Analysis

In this section, I begin with a discussion of the Hindi-Urdu phonemic inventory and of which phonemes are likely to be too rare to draw conclusions on. I then discuss the orthographic representations of the various classes of phonemes, starting with the consonants. For each phoneme for which I have sufficient data, I determine the possible orthographic representations and discuss what these tell us about linguistic, psychological, sociocultural, and technological fit in Roman Hindi-Urdu.

3.1. Phonemic Inventory

Table 1 displays the generally accepted phonemic consonant inventory of Hindi-Urdu (Ohala, 1994; Schmidt, 2003; Shapiro, 2003). Perso-Arabic and Sanskrit loans are noted.

Hindi-Urdu has a four way stop contrast for the four main places of articulation, and some speakers additionally have a voiceless uvular stop. Palatal affricates also exhibit a four way voicing contrast, and most native consonants also exhibit a phonemic length contrast word-medially before lax vowels. The geminate nasals, fricatives, and approximants do not occur frequently enough in my data to analyze in detail, but I do analyze the geminate stops as a possible model for how other geminates could potentially be written. /d/ and /t/ are in complementary distribution in native words, but /d/ can occur in additional environments in English loanwords and thus is generally considered its own phoneme (Schmidt, 2003). Among nasal stops, only /m m: n n:/ are widespread across speakers and across phonemic environments; /ɳ ɳ̄ ɳ̄/ occur contrastively only in formal Hindi registers, and only in homorganic nasal-stop clusters (Ohala, 1983, p. 5). /f z ʃ x ɣ h/ also occur only among a subset of speakers. Several authors consider /v/ to be an approximant because it is often realized as [w] (Kachru, 2006; Schmidt, 2003; Shapiro, 2003); and some authors transcribe /v/ with /v/, as well (Ohala, 1994). But Pierrehumbert and Nair (1996) find clear frication in [v], and identify a pattern of complementary variation with [w]. Additionally, /v:/ is never pronounced as an approximant (Ohala, 1983, pp. 6-7). Thus, I analyze /v/ with the fricatives in my analysis.

	Bilabial	Labio-dental	Dental	Alve-olar	Retroflex	Palatal	Velar	Uvu-lar	Glottal
Stops/ Affricates	p p: ^h p ^h ; b b: b ^h		t t: ^h t ^h ; d d: ^h d ^h ; d̪ d̪: ^h d̪ ^h ;		t̪ t̪: ^h t̪ ^h ; d̪ d̪: ^h d̪ ^h ;	tʃ tʃ: ^h tʃ ^h ; dʒ dʒ: ^h dʒ ^h ;	k k: ^h k ^h ; g g: ^h g ^h ;	q ¹	
Nasals	m m:	f ¹ v v:		n n: s s: z ¹	n ² ɳ	ɲ ɟ ¹			h ² ɦ
Fricatives						j			
Approx- imants									
Lateral									
Approx- imants									
Taps									

TABLE 1. The phonemic consonant inventory of Hindi-Urdu

¹ Loaned from Perso-Arabic.

² Loaned from Sanskrit.

³ Loaned from Sanskrit and Perso-Arabic.

A few different analyses of the Hindi-Urdu vowel system have been proposed. The first of two main points of contention is whether Hindi-Urdu has phonemic vowel length or a phonemic tenseness contrast. According to Schmidt (2003), /i: e: u: o:/ and /ɪ ɛ ʊ ɔ/ do differ in quality, but have nonetheless generally been analyzed as differing in length. The second point of contention is whether Hindi-Urdu has any diphthongs, and if so, how many. Most phonological grammars of Hindi Urdu posit that there are no diphthongs (Kachru, 2006; Ohala, 1983; 1994; Schmidt, 2003; Shapiro, 2003); by contrast, several recent studies have found between two and eighteen diphthongs in Hindi-Urdu, but these seem to differ significantly by dialect (Khurshid, Usman, and Javaid, 2003; Mishra and Bali, 2011; B. D. Sharma, 2018).

Assuming a tenseness contrast and no diphthongs results in the phonemic inventory shown in Figure 1. All but /æ/ occur in native words and exhibit a phonemic nasalization contrast. /æ/ is generally considered to be an additional phoneme, but it mainly occurs in loans from English, so I will not discuss it explicitly except where relevant to the orthographic representations of other vowels.

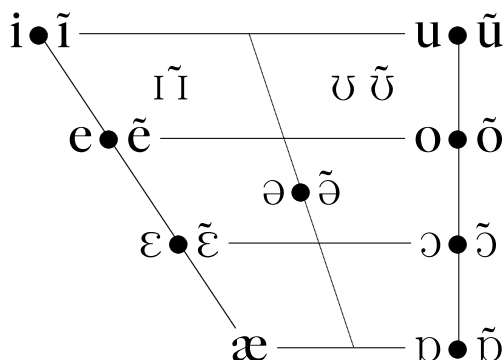


FIG. 1. An eleven vowel phonemic inventory of Hindi-Urdu (Ohala, 1983)

3.2. Oral Stops, Velar Fricatives, and Affricates

I begin by discussing the orthographic representations of the oral stops. Because most speakers have at least 30 oral stop phonemes in Hindi-Urdu, and the Roman script only has six standard letters for writing stops without diacritics, the stops are a good case for assessing the importance of linguistic fit. Yet, Roman Hindi-Urdu has developed remarkably consistent ways of marking aspirated, breathy voiced, and geminate

consonants. Place of articulation is represented with ⟨b d g p t k⟩, similarly to English: ⟨p b⟩ are bilabial stops; ⟨t d⟩ are dental stops; and ⟨k g⟩ are velar stops. ⟨c⟩ is in general not used: the only word in which it appears outside of ⟨ch⟩ is ⟨crore⟩ ‘ten million’, a spelling which dates back to 19th century British India (Oxford Dictionary Press, 1893). To represent gemination and voice quality contrasts, a second character is used. To represent an aspirated stop, ⟨h⟩ follows a voiceless stop symbol, and to represent a breathy voiced stop, ⟨h⟩ follows a voiced stop symbol. To represent a geminate stop, the stop symbol is repeated. The graphemic representations resulting from this system are shown in Table 2. Retroflex stops do not use unique symbols, and are instead represented using the same symbols as dental stops. I did not identify any aspirated retroflex stops, geminate retroflex stops, or aspirated geminate stops in the top 2000 words, although these are possible according to Ohala (1983, p. 2). If these follow the same pattern as the other stops, aspirated and geminate retroflex stops would be represented in the same way as their dental counterparts, and aspirated geminate stops could be represented using three-character graphemes with the stop character appearing twice, followed by ⟨h⟩. Further study will be required in order to determine whether this is the case.

TABLE 2. Representations of the various stop contrasts for voiced dental and retroflex stops

Place of Articulation	Bilabial	Dental	Velar	Retroflex
Singleton Unaspirated	⟨p⟩ ⟨b⟩	⟨t⟩ ⟨d⟩	⟨k⟩ ⟨g⟩	⟨t⟩ ⟨d⟩
Singleton Aspirated/Breathy	⟨ph⟩ ⟨bh⟩	⟨th⟩ ⟨dh⟩	⟨kh⟩ ⟨gh⟩	No data
Geminate Unaspirated	⟨pp⟩ ⟨bb⟩	⟨tt⟩ ⟨dd⟩	⟨kk⟩ ⟨gg⟩	No data

Although retroflex Hindi-Urdu stops occur in the same phonetic environments as dental stops, they are much rarer in native words. At the same time, retroflex stops occur almost exclusively in loanwords from English (Center for Language Engineering, n.d.). In the top 2000 words in my dataset, which excludes English loanwords, /t/ occurs 5 times and /d/ occurs twice, in all cases represented by ⟨t⟩ and ⟨d⟩. Some of these occurrences are listed in Table 3.

English loanwords often feature different phonotactic constraints from native words and they may also borrow English orthographic conventions. Thus, the distribution of retroflex stops in Hindi-Urdu may itself lead to relatively few cases of complete ambiguity. Additionally, ⟨t⟩ and ⟨d⟩ are convenient to type and may already be associated with the retroflex stops as a result of English orthography. If ambiguity between the retroflex and dental stops is indeed low, I hypothesize that using ⟨t d⟩ for both sets could maximize technological fit at low cost of psycho-

TABLE 3. Miscellaneous examples of ⟨t⟩ and ⟨d⟩

Roman spelling	⟨ghanta⟩	⟨theek⟩	⟨dal⟩	⟨tha⟩	⟨tum⟩	⟨desh⟩
Phonemic representation	[g ^h əntə]	[t ^h i:k]	[dɑ:l]	[t ^h ɑ]	[t̪um]	[de:ʃ]
Meaning	‘hour’	‘fine’	‘put’	‘was’	‘you’	‘country’
Occurrences	10	28	6	301	288	166

logical fit. The velar fricatives /x ɣ/ also reuse the stop representations in some cases. These sounds have no clear analogue in English orthography, and in Hindi-Urdu pedagogical literature and early colonial literature, these are typically represented by ⟨kh gh⟩ and differentiated from /k^h g^h/ only by diacritics. Complicating the situation is phonological variation: /x/ is frequently pronounced as /kh/ by some speakers, but /ɣ/ is often pronounced as [g] (Fatima, 2011). If phonological variation is represented in writing, then, one might expect /g^h/ to be represented by ⟨g⟩, but /k^h/ not to be represented by ⟨k⟩. Indeed, /ɣ/ is represented by ⟨g⟩ more frequently than by ⟨gh⟩, but /kh/ is exclusively represented by ⟨kh⟩. Full frequency counts for each orthographic representation are listed in Tables 4 and 5. Thus, I conclude that the representations of /x/ and /ɣ/ do reflect phonological variation.

TABLE 4. Frequency counts by phoneme represented by ⟨k⟩ and ⟨kh⟩

	⟨k⟩	⟨kh⟩
/k ^h /	0	1011
/x/	0	366
/k/	15558	0

TABLE 5. Frequency counts by phoneme represented by ⟨g⟩ and ⟨gh⟩

	⟨g⟩	⟨gh⟩
/g ^h /	0	178
/ɣ/	100	27
/g/	3326	0

Finally, I consider the voiceless uvular stop /q/. /q/ also has no obvious analogue in English orthography, and has historically been represented by ⟨q⟩ in early literature (Gilchrist, 1803; A. Rahman, 1923; R. N. Sharma, 1937). In similar fashion to /x ɣ/, /q/ surfaces as [k] for many speakers (Fatima, 2011; Shapiro, 2003). Like the retroflex stops, /q/ is relatively rare underlyingly, and I only identified 7 instances of it in the top 2000 most common words in the dataset. Table 6 shows the frequency counts of the letters used to represent /q/ in these words. On the one hand, the usage of ⟨k⟩ could reflect phonological variation, as was the case for /x ɣ/. But ⟨k⟩ could simply be another representation of /q/, which also seems plausible because of the rarity of /q/. Like the

retroflex stops, this would then be another possible case of orthographic representations being reused for rare phonemes.

TABLE 6. Orthographic representations of /q/ in the data

	⟨q⟩	⟨k⟩
/q/	129	66

Finally, the affricates /tʃ/ and /dʒ/ also exhibit an aspiration contrast as well as a length contrast, and in their singleton unaspirated forms are represented by ⟨ch⟩ and ⟨j⟩ respectively, which seem to be straightforwardly derived from English orthography. /dʒ/ shows the same tendencies as the stops, with ⟨jh⟩ and ⟨jj⟩ representing the breathy voiced and geminated forms, respectively. The fact that ⟨ch⟩ is two characters long complicates things, however. If /tʃ/ matched the behavior of the stops, ⟨chh⟩ would represent /tʃʰ/ and ⟨cch⟩ or ⟨chch⟩ would likely represent /tʃ:/. Instead, this pattern seems to break down, with /tʃ/ and /tʃʰ/ both being predominantly represented by ⟨ch⟩, and the rarer forms /tʃ:/ and /tʃʰ:/ having all four possible representations. Table 7 shows the full frequency counts. The instability of /tʃ:/ and /tʃʰ:/ indicates that three- and four-character representations may be harder to distinguish from each other; future work can test this hypothesis. If this is true, the rarity of longer orthographic representations could represent the maximization of psychological fit at the cost of linguistic fit.

TABLE 7. Orthographic representations of /tʃ tʃʰ tʃ:/ tʃʰ:/

	⟨cch⟩	⟨chh⟩	⟨ch⟩	⟨chch⟩
/tʃ/	0	0	600	0
/tʃʰ/	0	78	559	0
/tʃ:/	18	9	35	11
/tʃʰ:/	32	52	11	24

3.3. Nasal Stops

As with the oral stops, the orthographic representations of nasal stops are relatively straightforward, with possible ambiguities occurring only in particular phonetic environments among a relatively rare set of words. /n/ and /m/—the only nasal stops to occur in a wide range of

phonetic environments—are exclusively represented by ⟨n⟩ and ⟨m⟩, respectively. /ɳ ɳ ɳ/ are all rare in my data, and I selected words from the entire dataset in order to study these. /ɳ/ and /ɳ/ are represented by ⟨n⟩ in all of the examples in my data, and /ɳ/, which occurs rarely in my data, is written using ⟨y⟩ or ⟨n⟩. All three of these occur only before homorganic stops and affricates, and many speakers exhibit no contrast even in those cases. So ⟨n⟩ is ambiguous only in particular consonant clusters and only for certain speakers. Thus, as with retroflex stops, existing orthographic representations are again re-used to represent rare phonemes.

3.4. Sibilants and Fricatives

Next, I consider the sibilants. The main point of note here is that although the core Hindi-Urdu alveolar sibilants are /s ɖ̪/, /ʃ/ is present in Sanskrit and Perso-Arabic loans and /z/ is present in Perso-Arabic loans only. According to Ohala (1983, p. 4), most Hindi speakers pronounce /ʃ/ as [s] and /z/ as [ɖ̪̃]. In written form, I found several examples of /z/ being written with ⟨j⟩, but none of /ʃ/ being written with ⟨s⟩. The frequency counts for each grapheme used to represent /s z ʃ ɖ̪̃/ are shown in Tables 8 and 9. And finally, the Sanskrit loan /ʃ/ is rare in my data but appears to be represented by ⟨sh⟩ when it does appear. Again, rare phonemes seem more likely to re-use orthographic representations.

TABLE 8. Orthographic representations of /z/ and /ɖ̪̃/

	⟨z⟩	⟨j⟩
/z/	465	192
/ɖ̪̃/	0	3846

TABLE 9. Orthographic representations of /s/ and /ʃ/

	⟨s⟩	⟨sh⟩
/s/	5885	0
/ʃ/	0	1256

Shapiro (2003) indeed considers /ʃ/ to be a standard pronunciation, whereas /z/ almost certainly is not, so the use of ⟨sh⟩ to represent /ʃ/ may be a result of more speakers contrasting /s/ and /ʃ/ in speech. Additionally, perhaps the fact that /ʃ/ exists in both Sanskrit and Arabic loans neutralizes sociocultural factors which might make some Hindi speakers less likely to contrast /z/ and /ɖ̪̃/ in writing even if they are aware of the contrast.

/f/ is another Perso-Arabic loan which is often pronounced as /p/ in speech. However, Shapiro (ibid.) and Mangrio and Ali (2014) also note the existence of hypercorrection in both Urdu and Hindi speakers, where /p/ may be pronounced as [f] even in native words. In the data, the only

such word I found was /p^h_{Ir}/ ‘then’, which was spelled as ⟨fir⟩ and ⟨phir⟩ at nearly equal frequency. This word is the most cited example of this phenomenon and perhaps the most common one in speech, so this may well reflect actual pronunciation. Otherwise, ⟨p⟩ consistently represents /p/ and ⟨f⟩ consistently represents /f/.

The final fricative to consider is /v/, where a case of allophonic variation may give evidence for how a phoneme may be represented when the correspondence to the Roman script is not easily perceived. According to Pierrehumbert and Nair (1996), /v/ is realized as [w] in the onglide position of a syllable; that is, between a syllable-initial consonant and a vowel, or between an obstruent and a word boundary. They found [v] and [w] to be in free variation between a word-medial singleton consonant and a vowel; apparently because the singleton consonant can be analyzed either as a syllable onset or as the coda of the previous syllable. Furthermore, Grover’s (2016) perceptual study found that native Hindi speakers could not reliably distinguish hearing [v] and [w]. If this is true, we might expect ⟨v⟩ and ⟨w⟩ both to be used to represent ⟨w⟩ regardless of phonetic environment.

In my data, ⟨w⟩ in fact occurs more frequently than ⟨v⟩ in all positions, including those that are realized as [v] in speech. The first two rows of Table 10 show the frequency counts of ⟨v⟩ and ⟨w⟩ for phonetic environments where [v] is the expected surface form, and the third row shows the scenario where [v] and [w] are in free variation. I have insufficient data for words where [w] is the expected surface form, but given that ⟨w⟩ is already the more likely orthographic representation, ⟨w⟩ appears to be the more likely representation across the board. I conclude, then, that the orthographic representation of /v/ does not reflect its phonetic realization, but given a perceived ambiguity in sounds represented by ⟨v⟩ and ⟨w⟩, ⟨w⟩ may be organically developing to be a standard form, possibly reflecting a tendency toward higher linguistic fit.

TABLE 10. Orthographic representations of /v/ by word position

Phonetic Environment	Expected Surface Form	⟨v⟩	⟨w⟩
Word-initial	[v]	200	984
Word-medial, following a vowel	[v]	76	233
Word-medial, following a singleton consonant	[v] or [w]	46	107

3.5. Approximants

Hindi-Urdu has the lateral approximant /l/ and four rhotic approximants/ɽ ɽʱ ɽ ɽ:/ . Previous work generally analyzes all of these as taps or flaps, (Kachru, 2006; Ohala, 1994; Schmidt, 2003; Shapiro, 2003), but Ohala (1994) notes that /ɽ/ is often realized as a trill, and /ɽ:/ is always realized as a trill. The lateral approximant takes the straightforward representation ⟨l⟩ with little ambiguity, but the rhotic approximants are more complicated. I have already noted that retroflex *stops* are not distinguished orthographically from other apical stops. It is notable, then, that /ɽ/ is often written as ⟨d⟩. This distinguishes the retroflex and alveolar flaps from each other, but it means /d/, /d/, and /ɽ/ are predominantly written the same way. Since /d/ and /ɽ/ occur in complementary distribution except in English loanwords, there would be no ambiguity between these two phonemes for any of the words in my dataset. There could potentially be confusion in English loanwords, but since these often follow different phonotactic constraints, it is possible that they would be recognizable using other features.

The frequencies of all of the orthographic representations which represent rhotic approximants are shown in Table 11. A future perceptual study may shed light on whether distinguishing /ɽ/ from /ɽ/ turns out to be more meaningful with regard to reducing ambiguity.

TABLE 11. Orthographic representations of rhotic approximants

	⟨r⟩	⟨rr⟩	⟨d⟩	⟨dd⟩	⟨rh⟩	⟨dh⟩
/ɽ/ or /ɽ:/	6223	33	0	0	0	0
/ɽ/	63	0	129	0	0	0
/ɽʱ/	0	0	0	0	0	33
/d/ or /d/	0	0	3470	0	0	0
/d:/	0	0	0	95	0	0
/dʱ/	0	0	0	0	0	462

It appears that /ɽ/ and /ɽ:/ are both represented by ⟨r⟩. However, /ɽ/ in the interjection /əre:/ and the word /d̪ər/ ‘fear’ seems to be sometimes represented by ⟨rr⟩ (see Table 12), and /ɽ:/ seems to never be represented by ⟨rr⟩. The reason for this is unclear, but it is possible that a larger dataset could show more uses of ⟨rr⟩. Overall, possible correspondences between the different taps and Roman script representations are much less obvious, and this seems to lessen the likelihood of achieving high linguistic fit for these phonemes.

TABLE 12. Orthographic representations of /əre:/ ‘hey!’ and /dər/ ‘fear’

	/əre:/	/dər/
⟨rr⟩	23	10
⟨r⟩	32	16

3.6. /j/ and /ɸ/

The consonants that remain to be examined are /j/ and /ɸ/. Both of these are notable in my data because they allow me to consider whether Roman Hindi-Urdu inherits certain features from the historical Hindi and Urdu scripts.

/j/ often appears before or after a high or mid front vowel, as in [dija] ‘give’ and [gəji] ‘go’-3pst (ibid.). These are generally not transcribed as vowel hiatus—and the glide is written in both the Hindi and Urdu orthographies—but vowel hiatus would presumably be a possible way to analyze these sequences. Table 13 shows the representations of various sequences that could potentially be perceived either as vowel hiatus or as an approximant between two vowels. In all of these sequences, it is more common to include the ⟨y⟩ than to omit it. I conclude that these sequences are typically not viewed as vowel hiatus, and this could be either because of how they are perceived in speech or because of how they have historically been written.

TABLE 13. Representations of high vowel-glide-vowel clusters

Representation	⟨iya⟩	⟨ia⟩	⟨iye⟩	⟨ie⟩	⟨iyo⟩
Phonetic Form	/ijɑ/	/ijɑ/	/ije/	/ije/	/ijo/
Occurrences	378	64	88	51	13

A notable feature of /ɸ/ is that both the Hindi and Urdu orthographies sometimes mark a historical word-final /ɸ/ which is now deleted (Oberlies, 2005). Of course, a word-final orthographic ⟨h⟩ is also common in English in words such as *ah* and *oh*. So, I consider whether an orthographic word-final ⟨h⟩ exists in Roman Hindi-Urdu and whether it is more likely in words where this mirrors the Hindi and Urdu representations. The vast majority of vowel-final words in the data are written with no final ⟨h⟩. Of the words that are indeed written with word-final ⟨h⟩, the majority end in low vowels, and the majority have a word-final ⟨h⟩ in Hindi orthography, as shown in Table 14. The representations in Urdu orthography often overlap but may not match exactly. Still, this is

potentially another case of features of historical writing systems being carried over to the Roman script.

TABLE 14. ⟨h⟩-final Roman Hindi-Urdu words by their Hindi representation

Word-final /h/ represented in Hindi Orthography	No word-final /h/ represented in Hindi
566	352

3.7. Oral Vowels

Of the eleven vowels in the Hindi-Urdu phonemic inventory, all of them are present in several dialects of English. Yet, English orthography does not represent these uniquely, which makes the vowels an interesting study of how the Roman script maps onto Hindi-Urdu phonemes. Devanagari does represent these uniquely (Shapiro, 2003), whereas Perso-Arabic represents the vowels with a total of only three graphemes. In his Roman-script translations of Aesop's Fables, Gilchrist (1803) devises his own system for representing the vowels: He adds ⟨uo⟩, ⟨ee⟩, ⟨oo⟩, and ⟨y⟩, but still appears to use ⟨oo⟩ to represent both /u/ and /ʊ/. Pedagogical materials published since have used a wide variety of systems: A. Rahman (1923) and R. N. Sharma (1937) use ⟨ī ē ū⟩ to represent /i e u/ and ⟨au⟩ to represent /ɔ/; and Khan (2000) uses capital letters to represent the lax vowels.

I now analyze the orthographic representations of vowels in my data. I generally will focus on monophthongs in this analysis, but /ai oi au əu/—which may be diphthongs in some dialects, and cases of hiatus in others—appeared prominently in my data and their representations sometimes overlapped with the representations of monophthongs, so I include these as well in order to provide a fuller picture of the monophthongs.

First, I consider the front vowels as well as /ai/ and /oi/. Frequency counts of the graphemes and grapheme combinations used to represent these are shown in Table 15. ⟨i⟩ almost always represents both /i/ and /ɪ/. Using ⟨ee⟩ for /i/ is a possible way to draw that distinction, but is relatively uncommon. Interestingly, this is a departure from historical Roman-Urdu conventions at the expense of lower linguistic fit, and the reason for this is not entirely clear, given that ⟨ee⟩ is commonly used in English orthography to make exactly this distinction. A larger distinction is drawn between /e/ and /ɛ/, with ⟨e⟩ typically used to represent /e/, and ⟨ai⟩ typically used to represent /ɛ/. This, in turn, blurs the dis-

inction between /ai/ and /ε/, but since /ai/ is not necessarily monosyllabic, it is possible that that distinction is clearer from context. /oi/ occurs in only one of the 2000 most common words, /koi/ ‘somebody’, and it is generally represented by ⟨oi⟩.

TABLE 15. Orthographic representations of front vowels and potential diphthongs involving front vowels

	⟨i⟩	⟨ai⟩	⟨oi⟩	⟨ii⟩	⟨e⟩	⟨ee⟩	⟨oyi⟩
/ɪ/	1246	0	0	0	2	0	0
/i/	2244	9	0	6	44	342	0
/ε/	0	806	0	0	163	0	0
/ai/	0	3986	0	0	0	0	0
/oi/	0	0	378	0	0	0	11
/e/	14	1320	0	0	13570	6	0
/ə/	0	0	0	0	137	0	0

I now turn my attention to sequences involving back vowels. The frequency counts of their representations are shown in Table 16. Analogous to my findings for front vowels, ⟨oo⟩ seems to have been largely abandoned in favor of using ⟨u⟩ to represent both /ʊ/ and /u/. However, ⟨o⟩ is used to represent /o/ and ⟨au⟩ generally represents /ɔ/. Thus, for both front and back vowels, high vowels are more ambiguous in Roman Hindi-Urdu than mid-vowels, and this may be evidence that tenseness distinctions between the mid vowels are more salient to speakers. /a/ is represented by both ⟨a⟩ and ⟨aa⟩, but neither of these represents any other phonemes.

I observe considerable variation in the representation of /ə/. Shapiro (2003) considers this sound to be mid-central as in [ə] or [ʌ], whereas Ohala (1994) considers it to be lower, closer to [ɐ]. That would make ⟨a⟩, which represents other low vowels, a possible candidate. Orthographically, both the Perso-Arabic and Devanagari writing systems mark /ə/ only at the beginning of a word. Interestingly, Gilchrist (1803) uses ⟨u⟩

TABLE 16. Orthographic representations of back vowels

	⟨o⟩	⟨u⟩	⟨au⟩	⟨ou⟩	⟨oo⟩	⟨a⟩	⟨aa⟩
/ʊ/	0	3026	0	0	0	0	0
/u/	52	1368	0	0	225	0	0
/o/	6560	0	0	48	31	0	0
/ɔ/	239	0	927	0	0	0	0
/au/	0	0	14	0	0	0	0
/əu/	0	0	13	0	0	0	0
/ɑ/	0	0	0	0	0	3787	2961

to represent /ə/. Additionally, /ə/ is often fronted to [ɛ] before /h/, where the sound following /h/ is not /ə/ (Pandey, 1991). /ə/, then, has several contenders for possible orthographic representations. And as with /v/, /ə/ gives us another opportunity to see whether an allophonic process is reflected in Roman orthography. The orthographic representations of /ə/ in my data are shown in Table 17.

TABLE 17. Orthographic representations of /ə/

Representation	Ø	⟨a⟩	⟨e⟩	⟨i⟩	⟨u⟩
Occurrences	1140	12327	126	12	276

/ə/ is most commonly represented by ⟨a⟩, while using no vowel is the second most common strategy, and the use of ⟨e⟩ and ⟨i⟩ is rare. ⟨u⟩ seems to appear in words involving the first-person pronoun morpheme /hiəm/, but in no other words. Since /hiəm/ is a common morpheme, this may be a convention that was carried over from early publications. That there is so much variation may indicate that phonemic /ə/ is not consistently perceived as any particular vowel. It is possible that perception of /ə/ depends on its phonological environment; this would resemble English, where /ə/ typically takes on acoustic properties of neighboring vowels (Cole, Linebaugh, Munson, and McMurray, 2010). Since ⟨a⟩ is the most common representation, it may be the most common way that /ə/ is perceived, or this may be an example of linguistic fit being arbitrarily maximized over time. Future work can further examine whether the representations of /ə/ differ based on their phonetic environment.

Another observation from Table 18 is that the omission of /ə/ in the orthography is more common than the use of ⟨e⟩, ⟨i⟩, and ⟨u⟩, but is not close to being the preferred strategy. It is possible that doing this more often would introduce ambiguity; in the data, the majority of omissions of /ə/ indeed do not create ambiguity. The most frequent examples are shown in Table 18.

TABLE 18. The four most common words where /ə/ is represented by the absence of a vowel

Orthographic representation	⟨nhi⟩	⟨kr⟩	⟨rha⟩	⟨pr⟩
Phonemic form	[nəhi]	[kəɾ]	[rəfi]	[pəɾ]
Meaning	‘no’	‘do’	‘stay’	‘on’
Occurrences	348	154	93	76

3.8. Nasal Vowels

Nasal vowels are used in Hindi-Urdu morphology for inflecting both nouns and verbs. In both cases, nasalization of the final vowel, or the addition of a nasalized vowel to the end of the word, serves to mark pluralization of a noun or of a verb conjugation. Harbour (2021) points out that many languages do not mark such morphological processes phonetically, and even those that do often don't mark them orthographically. Both the Hindi and Urdu orthographies do mark nasalization however, and both do so by adding a separate character in addition to the vowel. In Perso-Arabic, this character is derived from the representation of ⟨n⟩, whereas Devanagari has a separate character specifically to indicate nasalization.

Outside of morphological processes, Ohala (1983) and Masica (1991) agree that Hindi-Urdu has phonemic nasal vowels word-finally and word medially before voiceless consonants. However, even when expanding my dataset to all words, I identified only four examples of word-medial nasal vowels. These are listed in Table 19 along with their frequencies. Although it is difficult to draw conclusions on such a small sample size, ⟨n⟩ is indeed used to distinguish word-medial nasal vowels from oral vowels. Note that use of the ⟨n⟩ matches the convention used in the Urdu writing system.

TABLE 19. Occurrences of underlying word-medial nasal vowels in the fully expanded dataset

Orthography	⟨saanp⟩	⟨ansu⟩/⟨aansu⟩	⟨hanso⟩
Meaning	‘snake’	‘tears’	‘laugh’
Phonetic form	[sā:p]	[ā:su]	[fāso]
Occurrences	2	6	1

Because phonemic nasal vowels are relatively rare other than in morphological inflections, I expand the dataset to the top 4000 most frequent words in my data. These include eight words with phonemic nasal vowels: /nəfī/ ‘no’, /fiā:/ ‘yes’, /kjū/ ‘why’, /jəhī/ ‘here’, /jafiā:/ ‘here’, /kafiā:/ ‘where’, /d̪ʒafiā:/ ‘where’, and /vafiā:/ ‘there’. Again, nasalization can be indicated by inserting ⟨n⟩ after the vowel; but the nasal vowels in these words are much more frequently written without marking nasalization. Table 20 shows the frequency counts of each of the possible forms.

Finally, I examine word-final nasal vowels that do result from inflectional processes. The most frequent of these processes is pluralization of both noun and verb forms, where a word-final oral vowel is nasalized, or a nasal vowel is added to a word-final consonant. In both cases, the plural form often forms a minimal pair with the singular form, which makes it

TABLE 20. Frequencies of the different representations of nasal vowels in the top 4000 words in the dataset, excluding nasal vowels resulting from inflectional processes. V represents a vowel grapheme.

Orthographic Form	V⟨n⟩	V
Frequency	289	1476

difficult to quantify cases where the nasalized form is not marked. Chief among these is [fĩe:], the third person plural form of ‘to be’, which often cannot be disambiguated from the singular form [fĩe:] without examining each post containing them. Table 21 lists the representations of these words in the data with their frequencies—the forms ending in ⟨n⟩ are unambiguously plural.

TABLE 21. Representations of the third person ‘to be’ forms

Representation	⟨hai⟩	⟨hain⟩	⟨h⟩	⟨hen⟩	⟨hein⟩	⟨hn⟩
Phonetic forms	[fĩe:] or [fĩe:]	[fĩe:]	[fĩe:] or [fĩe:]	[fĩe:]	[fĩe:]	[fĩe:]
Occurrences	3285	411	349	20	17	14

I now turn my attention to the remainder of the inflected forms. Once again expanding to the 4000 most frequent words, the word [betjã] ‘daughters’ is written with a word-final ⟨n⟩ 7 times, and without a word-final ⟨n⟩ 8 times. For this word, there is no ambiguity between the singular and plural forms. However, I identified 295 other inflected forms with word-final nasal vowels, and all of them were written with nasalization marked. The most frequent of these are shown with their frequencies in Table 22.

TABLE 22. Common morphological inflections with word-final nasal vowels

Representation	⟨logon⟩	⟨doston⟩	⟨walon⟩
Phonetic form	[lo:gõ]	[do:stõ]	[va:lõ]
Meaning	‘people’	‘friends’	‘ones’
Occurrences	75	28	21

I conclude, then, that nasalization is marked when it is a result of grammatical inflection, but is not marked otherwise. This contradicts Harbour’s (2021) finding which indicates that inflection is less likely to be marked orthographically. One possible reason is that nasalization is marked in both the Devanagari and Perso-Arabic writing systems, which

itself may have originated because of the existence of phonemic nasal vowels outside of grammatical inflection. When it comes to explaining why nasalization is often not marked, the most common instances of this are in very common words. Such words would be easier to recognize, and may also be more likely to have nonstandard forms develop for them which are easier to type. For instance, words which lack a nasalization marker are also more likely to lack some vowel representations, and some of these are presented with their frequency counts in Table 23.

TABLE 23. Miscellaneous abbreviated forms among words with nasal vowels

Representation	⟨nhi⟩	⟨nh⟩	⟨nhe⟩	⟨h⟩	⟨hn⟩
Phonetic forms	/nəfiː/	/nəfiː/	/nəfiː/	/fĩ/ or /fĩː/	/fĩː/
Meaning	‘no’	‘no’	‘no’	‘be’-3	‘be’-3pl
Occurrences	348	11	7	349	14

I have analyzed each of the phonemes in Hindi-Urdu which were available in my data. I have identified the degree of linguistic fit in each phoneme’s representations and have attempted to identify the origin of each representation.

4. Discussion

In this section, I begin with a summary of each phoneme’s possible orthographic representations in my dataset. I then discuss the broader patterns that seem to occur throughout my data, and summarize the factors that appear to affect orthographic representation as well as to what degree linguistic fit, psychological fit, and technological fit appear to be maximized.

Table 24 lists the orthographic representations I have identified for each phoneme which I analyzed. The phonemes are listed in the order in which they appear in Section 3, and the representations for each phoneme are listed in decreasing order of frequency. Representations that are markedly rare are marked as “(marginal)”.

TABLE 24. Correspondences between orthographic representations and phonemes

IPA	Roman	IPA	Roman
/p/	⟨p⟩	/n/	⟨n⟩
/b/	⟨b⟩	/ɳ/	⟨n⟩

Continued on next page

Table 24 – continued from previous page

IPA	Roman	IPA	Roman
/t̪/	⟨t⟩	/ɲ/	⟨n⟩
/d̪/	⟨d⟩	/ɲ/	⟨n⟩ or ⟨y⟩
/k/	⟨k⟩	/s/	⟨s⟩
/g/	⟨g⟩	/z/ (/d̪/ for some speakers)	⟨z⟩ or ⟨j⟩
/t/	⟨t⟩	/ʃ/ (/s/ for some speakers)	⟨sh⟩
/d/	⟨d⟩	/ʃ/	⟨sh⟩
/p ^h /	⟨ph⟩ or ⟨f⟩	/f/	⟨f⟩
/b ^h /	⟨bh⟩	/v/	⟨w⟩ or ⟨v⟩
/t̪ ^h /	⟨th⟩	/l/	⟨l⟩
/d̪ ^h /	⟨dh⟩	/r/	⟨r⟩ or ⟨rr⟩
/k ^h /	⟨kh⟩	/ɾ/	⟨r⟩ or ⟨d⟩
/g ^h /	⟨gh⟩	/t̪ ^h /	⟨rh⟩
/p:/	⟨pp⟩	/ɦ/	⟨h⟩
/b:/	⟨bb⟩	/j/	⟨y⟩
/t̪:/	⟨tt⟩	/ɪ/	⟨i⟩ or ⟨e⟩ (marginal)
/d̪:/	⟨dd⟩	/i/	⟨i⟩, ⟨ee⟩, ⟨e⟩ (marginal), ⟨ai⟩ (marginal), or ⟨ii⟩ (marginal)
/k:/	⟨kk⟩	/ɛ/	⟨ai⟩ or ⟨e⟩
/g:/	⟨gg⟩	/ai/	⟨ai⟩
/x/ ([k ^h] for some speakers)	⟨kh⟩	/oi/	⟨oi⟩ or ⟨oyi⟩ (marginal)
/ɣ/ ([g] for some speakers)	⟨g⟩ or ⟨gh⟩	/e/	⟨e⟩, ⟨ai⟩, ⟨i⟩, or ⟨ee⟩ (marginal)
/q/	⟨q⟩ or ⟨k⟩	/u/	⟨u⟩
/d̪/	⟨j⟩	/u/	⟨u⟩, ⟨oo⟩, or ⟨o⟩ (marginal)
/d̪ ^h /	⟨jh⟩	/o/	⟨o⟩, ⟨ou⟩ (marginal), or ⟨oo⟩ (marginal)
/d̪:/	⟨jj⟩	/ɔ/	⟨au⟩ or ⟨o⟩
/t̪f/	⟨ch⟩	/au/	⟨au⟩
/t̪f ^h /	⟨ch⟩ or ⟨chh⟩	/əu/	⟨au⟩
/t̪f:/	⟨ch⟩, ⟨cch⟩, ⟨chch⟩, or ⟨chh⟩	/a/	⟨a⟩ or ⟨aa⟩
/t̪f ^h :/	⟨chh⟩, ⟨cch⟩, ⟨chch⟩, or ⟨ch⟩	/ə/	⟨a⟩, , ⟨u⟩, ⟨e⟩, or ⟨i⟩ (marginal)
/m/	⟨m⟩	Ṽ	V or V⟨n⟩

In Table 25, I categorize each phonemic consonant's representations as having likely originated from either English orthography, historical Hindi-Urdu writing systems, colonial and early pedagogical literature, or the closest Roman script analogue according to users' auditory perception. Where applicable, typical orthographic conventions of English and other Roman scripts are generally followed. Orthographic representations are most consistent for sounds with analogues in English; these are /b d g p t̪ k tʃ dʒ m n f s ʃ/. The written forms of /ɦ/ and /j/ also fall into this category, but word-final ⟨h⟩ and intervocalic ⟨y⟩ reflect similar Devanagari and Perso-Arabic conventions. The use of ⟨h⟩ as an aspiration marker also reflects the way aspiration is represented in Perso-Arabic. The use of ⟨kh gh q⟩ to represent /x ɣ q/ can be traced back to early transliterations and pedagogical literature. The representations of /r ɽ v/ have no such history, so their use may reflect users' auditory perception, reduction of ambiguity with other phonemes, or simply an arbitrary development among comparably plausible phonemes. Neither Hindi and Urdu keyboard layouts nor the visual shape of the letters appear to have an influence on the orthographic conventions I observe.

TABLE 25. Hypothesized origins of the orthographic representations for each phonemic consonant

English orthography	Devanagari and Perso-Arabic writing systems	Early literature	Perception of sounds without clear equivalents
/b d g p t̪ k tʃ dʒ tʃ m n ɳ ɲ f s ʃ l/	/ɦ j/	/x ɣ q/	/r ɽ v ɳ/

Linguistic fit, technological fit, and psychological fit all appear to play a role in predicting preferred representations of phonemes. Technological fit is maximized by avoiding any diacritics, and the dispreference for three-character and longer representations may reflect such representations being harder to process. As for the linguistic fit of the various orthographic representations, Table 26 summarizes these for the consonants. Geminate consonants and aspirated stops are omitted for clarity, but the mechanisms to represent aspiration and gemination also lead to one-to-one correspondences. In general, higher linguistic fit is preferred. There is no starker example of this than the Roman representations of /k/ and /tʃ/: in English, /k/ is represented by both ⟨c⟩ and ⟨k⟩; and in both historical and modern Roman Hindi-Urdu publications, /tʃ/ is represented by both ⟨ch⟩ and ⟨c⟩. In my data, however, ⟨c⟩ virtually never appears outside of ⟨ch⟩. /v/ provides another example of the possible tendency towards a one-to-one correspondence, where ⟨w⟩ is

more commonly used than ⟨v⟩ even though speakers perceive /v/ as [w] and [v] at approximately equal rates. The representations of /ɔ/ and /ε/ display this tendency as well.

TABLE 26. Linguistic fit of the Roman orthographic representations of Hindi-Urdu consonants

One-to-one correspondence	Phonemes represented in multiple ways	Orthographic representations used for multiple phonemes
/b ɖ ɡ p t̪ k d̪ t̪ m ɳ ɳ f s ʃ l/	/v z t̪ʰ q x ɣ /	⟨d dd g kh r d t n y sh⟩

A few factors do appear to result in lower linguistic fit. These are summarized in Table 27. The first of these is the absence of a phoneme in English, or the lack of an obvious choice in Roman script. Hindi-Urdu /x ɣ q/ do not have phonemic analogues in English, and /v/ has two equally likely phonemic analogues according to auditory perception. This helps to explain the instability of the representations of each. The second possible factor is dialectal variation in how various phonemes are pronounced in speech. That /ɣ q z/ all have multiple orthographic representations which represent known phonological variation—and /x/ does not have multiple representations, because it is coincidentally represented identically to its alternative surface form /kʰ/—is evidence for this. Third, the lack of overlap in the phonetic environments of two graphemes seems to lessen the cost with regard to linguistic and psychological fit of using single graphemes to represent multiple phonemes. The retroflex consonants, which are generally written the same way as their non-retroflex counterparts, are an example of this, where the convenience of a more intuitive representation may override the disruption of linguistic and psychological fit. /ɳ/ and /ɳ/ exhibit the same pattern. Fifth, linguistic fit seems to be harder to maintain for longer orthographic representations. /t̪:/, /t̪ʰ/, and /t̪ʰ:/, all of which are often represented by three and four character combinations, display more variation in orthographic form than those geminates and aspirated stops which are written with one- and two-character Roman representations.

The high degree of variation in phonetic realization of vowels makes it difficult to determine whether differences in representation signify dialectal differences or differences in orthographic convention. Additionally, several Hindi-Urdu vowels may have phonetic realizations that resemble other vowels (Schmidt, 2003), which further complicates the analysis. The biggest difference I observe in vowels is that in multiple cases, the orthography does not take advantage of every potential orthographic representation to maximize linguistic fit. The lax/tense dis-

TABLE 27. Considerations that lead linguistic fit to decrease

Consideration	Example phonemes
Lack of an obvious choice in Roman script	/x ʏ q v r ʈ/
Dialectal variation	/p x ʏ q z/
Rarity of phonemes	/t d ʂ ɳ q/
Perceptual similarity to other phonemes	/t d r ʈ/
Length of the orthographic representation	/tʃ: tʃʰ/

inction is generally not reflected in the orthography for high vowels, despite the availability of ⟨ee⟩ and ⟨oo⟩; yet, ⟨au⟩ and ⟨ai⟩ are used in order to make the lax/tense distinction visible for mid vowels. This may reflect differences in the salience of that contrast for different vowel heights. Finally, the multitude of possible orthographic representations of /ə/ may reflect differences in the perception of that vowel in different phonetic environments, as well as the apparent tendency not to omit vowels in Roman orthography.

I have shown that Roman Hindi-Urdu reflects phonological variation between speakers, and some of these cases of variation are similar to cases of sociocultural variation found in spoken language, such as the use of phonemes loaned from Perso-Arabic (*ibid.*). I do not have the demographic data to determine whether any variation is the result of sociocultural factors, but it is plausible that users could intentionally choose from among these forms to identify themselves with or distance themselves from regional or religious groups associated with particular forms. The representations of /q/, /x/, /ʏ/, /z/, and /p/ all could reflect this phenomenon. Another potential way of signaling sociocultural identity in the Roman script is to carry over written conventions from the Urdu and Hindi orthographies, since these are already associated with speakers' sociocultural identity. Orthographic representations of vowels are one arena in which this could appear. Further work which takes users' demographic information into account can test these hypotheses directly and also may shed light on other sources of sociocultural variation.

5. Future Directions

Many future studies could expand upon this work. With regard to determining the orthographic representations of each phoneme, the rarer phonemes could be studied much more thoroughly given more data. Additionally, more data could allow for more detailed study of how phonetic and orthographic environments might affect the representation of

phonemes; for instance, how particular consonant clusters are written, or whether the representations of vowels such as /ə/ depend on their phonetic environments.

Several findings in this study indicate that psychological fit affects the representations of some phonemes, and psycholinguistic studies could test whether graphemes which represent multiple phonemes or long representations take longer to read and write, and whether the rarity of a phoneme affects processing speed. Similarly, much work remains to be done on sociocultural fit: future work could compare the orthographic conventions used by self-identified Urdu and Hindi speakers, or of internet users in India and Pakistan. Finally, surveys could be conducted to determine whether users have a conscious philosophy for how to write Hindi-Urdu in the Roman script, as well as to learn about users' intuitions as to which representations look 'correct' to them.

6. Conclusion

This study uses data from X to document the orthographic conventions of Hindi-Urdu and discusses the possible factors which have shaped these conventions in the absence of prescriptive standardization. The Roman script has become one of the main ways of using Hindi-Urdu in written form, and this study can help us understand how it is used. Linguistic fit, historical orthographic conventions, and users' auditory perception of phonemes all appear to help shape the orthographic conventions of Roman Hindi-Urdu, and they all might be expected to play a role in other organically developing writing systems. Finally, I have discussed some of the questions which remain to be answered and how future work may address them.

References

- Ahmad, R. (2011). "Urdu in Devanagari: Shifting orthographic practices and Muslim identity in Delhi". In: *Language in Society* 40.3, pp. 259–284.
- Akbari, M. (2013). "A preliminary linguistic analysis of Romanized Persian SMS messages". In: *Journal of Novel Applied Sciences* 2.8, pp. 197–205.
- Androutsopoulos, J. (2009). "'Greeklish': Transliteration practice and discourse in the context of computer-mediated digraphia". In: *Standard Languages and Language Standards: Greek, Past and Present*. Ed. by A. Georgakopoulou and M. Silk. Farnham: Ashgate.
- Atta, A. (2021). "Scripts on linguistic landscapes: A marker of hybrid identity in urban areas of Pakistan". In: *Journal of Nusantara Studies (JONUS)* 6.2, pp. 58–96.

- Bali, K. et al. (2014). ““I am borrowing *ya* mixing?”: an analysis of English-Hindi code mixing in Facebook”. In: *Proceedings of the first workshop on computational approaches to code switching*, pp. 116–126.
- Center for Language Engineering (n.d.). *Urdu 5000 Most Frequently Used Words List*. https://cle.org.pk/software/ling_resources/UrduHighFreqWords.htm.
- Cole, J. et al. (2010). “Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach”. In: *Journal of phonetics* 38.2, pp. 167–184.
- Faruqi, S. R. (2003). “A Long History of Urdu Literary Culture, Part I”. In: *Literary cultures in history: Reconstructions from South Asia*. Ed. by S. Pollock. University of California Press, pp. 805–63.
- Fatima, S. (2011). “Phonological Variation in Perso-Arabic Words in Urdu”. In: *Language in India* 11.7.
- Gilchrist, J. B. (1803). *Polyglot fables*. printed at the Hurkaru Office.
- Grover, V. (2016). “Perception and Production of /v/ and /w/ in Hindi speakers”. PhD thesis. City University of New York.
- Harbour, Daniel (2021). “Grammar Drives Writing System Evolution”. In: *Grapholinguistics in the 21st Century 2020. Proceedings: Grapholinguistics and Its Applications*. Ed. by Y. Haralambous. Fluxus Editions, pp. 201–221.
- Kachru, Y. (2006). *Hindi*. John Benjamins Publishing.
- Kanoongo, U. (2016). “Sociolinguistic Functions of Roman-Romanagari Code-switching in WhatsApp Instant Messaging”. In: *The Journal of Rase* 14.28, p. 14.
- Khan, M. Ahmad (2000). *Urdu phonology*. Dept. of Linguistics, Aligarh Muslim University.
- Khurshid, K., S. A. Usman, and N. Javaid (2003). “Possibility of Existence and Identification of Diphthongs and Triphthongs in Urdu Language”. In: *Crulp Annual Student Report*. Center for Language Engineering.
- Mangrio, R. A. and G. Ali (2014). “Rise and Fall of Arabic Loan Phonemes in Urdu”. In: *Literary Globalization and the Politics and Poetics of Translations* 15.
- Masica, C. P. (1991). *The Indo-Aryan languages*. Cambridge University Press.
- Meletis, D. (2018). “What is natural in writing?: Prolegomena to a Natural Grapholinguistics”. In: *Written Language & Literacy* 21.1, pp. 52–88.
- Meletis, D. and C. Dürscheid (2022). *Writing Systems and Their Use*. De Gruyter Mouton.
- Meyer, C. M. and I. Gurevych (2012). “Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography”. In: *Electronic Lexicography*. Ed. by Granger and Paquot. Oxford University Press, pp. 259–291.
- Mishra, D. and K. Bali (2011). “A Comparative Phonological Study of the Dialects of Hindi”. In: *ICPhS* 17, pp. 17–21.

- Mouresioti, Evgenia and Marina Terkourafi (2021). “Καλημέρα, kalimera or kalhmera?: A mixed methods study of Greek native speakers’ attitudes to using the Greek and Roman scripts in emails and SMS”. In: *Journal of Greek Linguistics* 21.2, pp. 224–262.
- Mukherjee, J. (2010). “The development of the English language in India.” In: *The Routledge Handbook of World Englishes*. Ed. by A. Kirkpatrick. Routledge, pp. 167–180.
- National Assembly of Pakistan (1973). *The Constitution of the Islamic Republic of Pakistan*. https://na.gov.pk/uploads/documents/1333523681_951.pdf.
- Nema, N. and J. K. Chawla (2018). “The Dialectics of Hinglish: A Perspective”. In: *Applied Linguistics Papers* 25.2, pp. 37–51.
- Noruzi, A. (2023). “Wiktionary Popularity from a Citation Analysis Point of View”. In: *Informology* 2.1, pp. 1–6.
- Oberlies, T. (2005). *A historical grammar of Hindi*. Leykam.
- Ohala, M. (1983). *Aspects of Hindi phonology*. Motilal Banarsidass Publishers Pvt. Limited.
- (1994). “Hindi”. In: *Journal of the International Phonetic Association* 24.1, pp. 35–38.
- Oxford Dictionary Press (1893). “Crore”. In: *Oxford English Dictionary*. <https://doi.org/10.1093/OED/8116865719>. DOI: 10.1093/OED/8116865719.
- Pandey, Pramod Kumar (1991). “Schwa fronting in Hindi”. In: *Studies in the Linguistic Sciences* 21.1.
- Parliament of India (1963). *The Official Languages Act*. en. https://www.indiacode.nic.in/bitstream/123456789/1526/1/A1963__19.pdf.
- Pierrehumbert, J. and R. Nair (1996). “Implications of Hindi prosodic structure. Current trends in phonology”. In: *Models and methods* 2, pp. 549–584.
- Putra, R. A. and S. Triyono (2018). “Outlandish spelling system invented by Indonesian internet society: The case of language usage in social networking site”. In: *International Journal of Applied Linguistics and English Literature* 7.7, pp. 66–73.
- Rahman, A. (1923). *Urdu conversational exercises*. Karachi: Azizi’s Oriental Book Depot.
- Rahman, T. (2011). *From Hindi to Urdu: a social and political history*. Oxford University Press.
- (2020). “Pakistani English”. In: *The handbook of Asian englishes*. Ed. by K. Bolton, W. Botha, and A. Kirkpatrick. Wiley-Blackwell, pp. 279–296.
- Ramanathan (2016). “English Education Policy in India”. In: *English Language Education Policy in Asia*. Ed. by R. Kirkpatrick. Springer.
- Rao, C. et al. (2011). “Orthographic characteristics speed Hindi word naming but slow Urdu naming: Evidence from Hindi/Urdu biliterates”. In: *Reading and Writing* 24.6, pp. 679–695.

- Rosowsky, A. (2010). "Writing it in English': script choices among young multilingual muslims in the UK". In: *Journal of Multilingual and Multicultural Development* 31.2, pp. 163–179.
- Salomon, R. (2003). "Writing systems of the Indo-Aryan languages". In: *The Indo-Aryan languages*. Ed. by G. Cardona and D. Jain. Routledge, pp. 75–114.
- Schmidt, R. L. (2003). "Urdu". In: *The Indo-Aryan languages*. Ed. by G. Cardona and D. Jain. Routledge, pp. 286–350.
- Sebba, M. (2007). "Orthography as social practice". In: *Spelling and Society*. Ed. by M. Sebba. Cambridge University Press, pp. 26–57.
- Shapiro, M. C. (2003). "Hindi". In: *The Indo-Aryan languages*. Ed. by G. Cardona and D. Jain. Routledge, pp. 250–285.
- Sharma, B. D. (2018). "Vowel Phonemes in Hindi". In: *East European Journal of Psycholinguistics* 5.2, pp. 71–91.
- Sharma, R. N. (1937). *Roman Urdu: A comprehensive study in Hindustani*. R.N. Sharma.
- Speer, R. (2022). *rspeer/wordfreq: v3.0 (v3.0.2)*. Version v3.0.2. DOI: 10.5281/zenodo.7199437.
- Unseth, P. (2005). "Sociolinguistic parallels between choosing scripts and languages". In: *Written language & literacy* 8.1, pp. 19–42.
- Wiktionary (2023). *Wiktionary, the free dictionary*. URL: https://en.wiktionary.org/wiki/Wiktionary:Main_Page.
- Zaidi, Syeda Bushra and Sajida Zaki (2017). "English language in Pakistan: Expansion and resulting implications". In: *Journal of Education & Social Sciences* 5.1, pp. 52–67.

