

GRAPHOLINGUISTICS AND ITS APPLICATIONS

Graphemics in the 21st Century

/ɣʁafematik/
Proceedings

Brest, June 13-15, 2018
Yannis Haralambous (Ed.)



Grapholinguistics and Its Applications 1

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 Graz, Austria*

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, *Sorbonne, Paris*

Jürgen Spitzmüller, *Universität Wien, Austria*

Richard Sproat, *Google, USA*

Susanne Wehde, *MRC Managing Research GmbH, Germany*

Yannis Haralambous (Ed.)

Graphemics in the 21st Century

/gɾafematik/

Brest, June 13–15, 2018

Proceedings

Fluxus Editions

Yannis Haralambous (Ed.). 2019. *Graphemics in the 21st Century. Brest, June 13–15, 2018. Proceedings* (Grapholinguistics and Its Applications, Vol. 1). Brest: Fluxus Editions.

This title can be downloaded at:

<http://fluxus-editions.fr/gla1.php>

© 2019, The respective authors

Published under the Creative Commons Attribution 4.0 License

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

ISBN: 978-2-9570549-0-9

e-ISBN: 978-2-9570549-1-6

ISSN: 2534-5192

DOI: <https://doi.org/10.36824/2018-graf>

Cover illustration: *Red PostScript Square and Gate Character*, using the painting “関 Kan (a frontier pass or gate) 1964, カーボン・水溶性ボンド、紙 carbon black and water-soluble glue on paper” by 井上有一 INOUE Yuichi (1916–1985), © UNAC TOKYO, permanent collection of Niigata City Art Museum, Niigata City, Japan. With permission by UNAC Tokyo.

Cover design and typesetting: Atelier Fluxus Virus

Main fonts: William Pro by Typotheque Type Foundry, Computer

Modern Typewriter by Donald E. Knuth, Source Han Serif

by Adobe Systems, Amiri by Khaled Hosny

Typesetting tools: X_YL^AT_EX, biblatex+biber (authoryear-icomp style),

xindex, titlecaseconverter.com

Fluxus Editions

38 rue Émile Zola

29200 Brest, France

www.fluxus-editions.fr

Dépôt légal : novembre 2019

Table of Contents

<i>Preface</i>	VII
<i>List of Participants at the Graphemics in the 21st Century 2018 Conference</i>	XI
FLORIAN COULMAS. – „Die Buchstabenschrift ist an und für sich die intelligentere.“ Überlegungen zur Bewertung von Schriftsystemen [“Alphabetic writing is in and for itself the more intelligent Form”. Reflections on the evaluation of writing systems]	1
MARC WILHELM KÜSTER. – Open and Closed Writing Systems. Some Reflections	17
MARTIN EVERTZ. – The History of the Graphematic Foot in English and German	27
YANNIS HARALAMBOUS & JOSEPH DICHY. – Graphemic Methods for Gender-Neutral Writing	41
WANG YIFAN. – What Are We Calling “Latin Script”? Name and Reality in the Grammatological Terminology	91
SVEVA ELTI DI RODEANO. – Digraphia: the Story of a Sociolinguistic Term	111
YANNIS HARALAMBOUS & MARTIN DÜRST. – Unicode from a Linguistic Point of View	127
CHRISTA DÜRSCHIED & DIMITRIOS MELETIS. – Emojis: A Grapho-linguistic Approach	167
KEISUKE HONDA. – What Do Kanji Graphs Represent in the Current Japanese Writing system? Towards a Unified Model of Kanji as Written Signs	185

TEREZA SLAMĚNÍKOVÁ. – On the Nature of Unmotivated Components in Modern Chinese Characters	209
CORNELIA SCHINDELIN. – The Li-Variation (隶变/隸變) <i>libiàn</i> . When the Ancient Chinese Writing Changed to Modern Chinese Script	227
KAMAL MANSOUR. – On the Origin of Arabic Script	245
JOSEPH DICHY. – On the Writing System of Arabic: The Semio-graphic Principle as Reflected in Nashī Letter Shapes	257
RAY STEGEMAN. – Orthographies in Papua New Guinea through the Years	269
DAVID ROBERTS, DANA BASNIGHT-BROWN & VALENTIN VYDRIN. – Marking Tone with Punctuation: Orthography Experimentation and Reform in Eastern Dan (Côte d’Ivoire)	293
KAVYA MANOHAR & SANTHOSH THOTTINGAL. – Malayalam Orthographic Reforms. Impact on Language and Popular Culture	329
NICOLAS BALLIER, ERIN PACQUETET & TAYLOR ARNOLD. – Investigating Keylogs as Time-Stamped Graphemics	353
PATRICIA THAINE & GERALD PENN. – Vocalic and Consonantal Grapheme Classification through Spectral Decomposition	367
<i>Index</i>	387

Preface

This volume of the Series *Grapholinguistics* gathers contributions by the participants of the *Graphemics in the 21st Century* (/grafematik/) conference that was organized by Yannis Haralambous with the support of IMT Atlantique and the CNRS (UMR 6285 LabSTICC, unit DECIDE) and was held in Brest from June 13 to June 15, 2018.

Its aim was to bring together disciplines concerned with writing systems and their representation in written communication, as well as to reflect on the current state of research in the area, and on the role that writing and writing systems play in neighboring disciplines like computer science and information technology, communication, typography, psychology, and pedagogy.

Not surprisingly, the papers gathered in this volume belong to various disciplines and consider writing from different points of view, involving linguistics, history, archeology, education, and natural language processing.

In his paper “‘Alphabetic writing is in and for itself the more intelligent Form’. Reflections on the evaluation of writing systems” (in German language), Coulmas takes Hegel’s statement in favor of the alphabetic script as a starting point and questions whether, and how, writing systems can be compared. This paper is followed by Küster’s “Open and Closed Writing Systems. Some Reflections,” which gives a different classification approach of scripts based on their openness, e.g., their possibility to increase their set of graphs.

The paper “The History of the Graphematic Foot in English and German” by Evertz deals with a notion inspired by the phonological foot unit (hierarchically located between the syllable and the word), the graphematic foot. The author discusses its pertinence, providing examples in English and German.

In several languages there have been attempts to provide gender-neutral forms. The paper “Graphemic Methods for Gender-Neutral Writing” by Haralambous & Dichy describes a particular case of gender-neutral forms, namely graphemic ones, in French, German, Greek, Italian, Por-

tuguese and Spanish. In the paper “What Are We Calling ‘Latin Script’? Name and Reality in the Grammatological Terminology,” Wang notices that the 81 languages using the Latin alphabet have only two letters in common, namely <A> and <I>, and raises the legitimate question whether the commonly-used term “Latin script” makes sense.

Elti di Rodeano, in her paper “Digraphia: the Story of a Sociolinguistic Term,” investigates the origin and evolution of the term “digraphia,” and gives a synthetic new definition.

The following two papers deal with Unicode encoding: in “Unicode from a Linguistic Point of View,” Haralambous & Duerst compare the basic Unicode technical terms (such as “character,” “character class,” and “character string”) with grapholinguistic notions and discuss ligatures and emojis. In “Emojis: A Grapholinguistic Approach,” Dürscheid & Meletis discuss emojis in more depth and raise questions about the role of the Unicode Consortium in present and future.

Three papers deal with the Chinese script. First of all, Honda’s “What Do Kanji Graphs Represent in the Current Japanese Writing system? Towards a Unified Model of Kanji as Written Signs,” discusses two approaches to Japanese kanji: the morphographic approach that states that they mainly represent morphemes, and the morphophonetic approach that states that they are primarily phonographic and only secondarily morphemic. After a discussion of these approaches, the author proposes a new model, combining them.

Slaměńíková, in “On the Nature of Unmotivated Components in Modern Chinese Characters,” takes the dual model of Chinese character component classification (semantic vs. phonetic) as a starting point, but investigates the role and linguistic importance of a third class of components which are neither semantic nor phonetic, and which she calls *unmotivated*. Schindelin, on the other hand, delves into the history of the Chinese script in her paper “The Li-Variation (隶变/隸變) *libiàn*. When the Ancient Chinese Writing Changed to Modern Chinese Script,” and sheds light on an historical turning point that occurred in the 1st century AD and changed the relationship between characters and components so that the script evolved into what it is today.

The next two papers deal with the Arabic script. Mansour, in “On the Origin of Arabic Script” contributes to the scholarly debate regarding the origin of the Arabic script, namely whether it derives from the Syriac script or rather from the Nabatean script. He concludes that both scripts have strongly influenced it, and that it is pointless to search for a unique ancestor. Dichy, in “On the Writing System of Arabic: The Semiographic Principle as Reflected in Nashī Letter Shapes,” discusses the interaction between Arabic grammar and Arabic script (in the Nashī style), both analytically and historically.

The following two papers deal with the activities of SIL International members in creating orthographies for unwritten languages. The

first one, Stegeman's "Orthographies in Papua New Guinea through the Years," gives an account of the various attempts over the past 60 years to provide graphemes for over 300 languages of all possible types. He describes the shift from linguistic zeal to underdifferentiation, motivated also by the new communication media. The following paper, "Marking Tone with Punctuation: Orthography Experimentation and Reform in Eastern Dan (Côte d'Ivoire)" by Roberts, Basnight-Brown & Vydrin, deals with a different part of the world: Côte d'Ivoire, where an experiment has been performed to evaluate a planned orthography reform (the previous orthographic system dating from 1982). The evaluation criteria have been the maximum ease of learning, of transfer and of reproduction.

The paper "Malayalam Orthographic Reforms. Impact on Language and Popular Culture" by Manohar & Thottingal describes an orthographic reform of the Malayalam script, introduced in 1971 by the State of Kerala, which affected ligatures representing the absence of vowel between consonants. As these ligatures are handled not on the Unicode character level but on the glyph level (by font technologies such as OpenType), by using the appropriate fonts it was possible to writers of Malayalam to return to the original script, as used before the reform, and the paper illustrates this tendency in the last years.

Finally, the two last papers of the volume deal with applications of graphemics in Natural Language Processing. The first one, by Ballier, Pacquetet & Arnold, titled "Investigating Keylogs as Time-Stamped Graphemics," considers keys of a keyboard representing graphemes and shows the relationship between keyboarding speed (provided in key-stroke logging datasets) and linguistic structure of the keyboarded text. The second paper, "Vocalic and Consonantal Grapheme Classification through Spectral Decomposition," by Thaine & Penn, also extract linguistic information from data, but this time the data are simple texts in a given (alphabetic or abjad) language and the linguistic information obtained is a classification of the characters into two classes, surprisingly similar to the classification into vowels and consonants as given in official grammars. The method used, namely *spectral decomposition*, is purely frequential and uses no external linguistic data.

All presentations at the *Graphemics in the 21st Century 2018* conference have been recorded and can be viewed on Youtube (https://www.youtube.com/playlist?list=PLJABkUSif0d8APXr0aZ2N96B5p2_0pAq9).

List of Participants at the *Graphemics in the 21st Century* 2018 Conference

Vlad Atanasiu, *University of Fribourg, Switzerland*
Nicolas Ballier, *University of Paris Diderot / CLILLAC-ARP, France*
Marc Bernot, *Thalès, Brest, France*
Nicolas Bouilleaud, *Codeurs en Liberté, Paris, France*
David Březina, *University of Reading, Rosetta Type Foundry, UK*
Florian Coulmas, *University of Duisburg-Essen, Germany*
Sveva Elti di Rodeano, *Università degli Studi di Udine, Italy*
Joseph Dichy, *Lyon – Aradic Editions, France*
Christa Dürscheid, *University of Zurich, Switzerland*
Martin J. Dürst, *Aoyama Gakuin University, Japan*
Martin Evertz, *University of Cologne, Germany*
Lina Fahed, *IMT Atlantique, Brest, France*
Yannis Haralambous, *IMT Atlantique, Brest, France*
Keisuke Honda, *Imperial College London, UK*
Cheikh Hito Kacfeh Emani, *IMT Atlantique, Brest, France*
Jan Kučera, *Charles University, Prague, Czech Republic*
Marc Wihelm Küster, *Budabe, Luxembourg*
Karvya Manohar, *Swathanthra Malayalam Computing, India*
Kamal Mansour, *Monotype, San Francisco, USA*
Grégoire Novel, *Ligature SASU, Paris, France*
Gerald Penn, *University of Toronto, Canada*
Morgane Pierson, *Atelier National de Recherche Typographique, Lyon, France*
Martin Raymond, *SIL International, Dallas TX, USA*
David Roberts, *Independent Researcher*
Cornelia Schindelin, *Johannes Gutenberg-Universität Mainz, Germany*
Tereza Slaměňíková, *Palacký University, Olomouc, Czech Republic*
Ray Stegerman, *SIL International, Dallas TX, USA*
Santosh Thottingal, *Swathanthra Malayalam Computing, India*
Yivan Wang, *The University of Tokyo, Japan*

„Die Buchstabenschrift ist an und für sich die intelligentere.“

Überlegungen zur Bewertung von Schriftsystemen

Florian Coulmas

Zusammenfassung. Sind manche Schriftsysteme besser (intelligenter) als andere? Gibt es sinnvolle und verlässliche Kriterien zur Bewertung von Schriftsystemen? Vor dem Hintergrund dieser Fragen vergleicht dieser Beitrag verschiedene Schriften. Ausgangspunkt ist das in der Überschrift zitierte Verdikt des deutschen Philosophen Georg Wilhelm Friedrich Hegel, das eine positive Beantwortung der Frage nach einer möglichen qualitativen Bewertung von Schriftsystemen nahelegt. Kritisch zu untersuchen sind jedoch die Maßstäbe, die angelegt wurden, um zu dem darin ausgedrückten Urteil zu kommen. Einer der zu diskutierenden Maßstäbe ist das Verhältnis von Schrift und Sprache, das in diesem Zusammenhang näher betrachtet werden soll.

“Alphabetic writing is in and for itself the more intelligent Form”. Reflections on the evaluation of writing systems

Abstract. Are some writing systems better (more intelligent) than others? Are there any criteria for the evaluation of writing systems that are both reasonable and robust? Against the background of these questions, this paper looks at several different writing systems. The point of departure is the verdict by German philosopher George Wilhelm Friedrich Hegel quoted in the title of the paper, which suggests that a qualitative evaluation of writing systems is possible. However, the yardsticks that were applied in order to arrive at the judgement expressed in it are to be critically examined. An important factor to be taken into consideration in this connection is the relationship between writing and language.

Schrift ist eine Technologie, die „Technologie des Geistes“, wie der Anthropologe Jack Goody (1977, S. 151) sie treffend nannte. Ein Leben ohne diese Technologie ist nicht mehr vorstellbar, auch wenn es auf diesem Planeten noch immer Menschen gibt, die sie nicht beherrschen. Nach heutigem Kenntnisstand kam es fünfmal unabhängig voneinander zu dieser Erfindung: im Indus, im Zweistromland, am Nil in Ägypten, am

Florian Coulmas
Institute of East Asian Studies and Faculty of Social Sciences
LE 645 - Forsthausweg 2, 47057 Duisburg, Germany
florian.coulmas@uni-due.de

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2534-5192), Vol. 1.
Fluxus Editions, Brest, 2019, p. 1–16. <https://doi.org/10.36824/2018-graf-coul>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

Gelben Fluss in China, und in Mittelamerika. Komplexere Gesellschaften und Formen der Zivilisation mit vielstufigen Verhältnissen der Produktion, des Austauschs und der Reproduktion erfordern Mittel nicht nur zur unmittelbaren Kommunikation, sondern auch zur senderunabhängigen Bewahrung von Information. Aus dieser Notwendigkeit entstand Schrift, besser: entstanden Schriftsysteme. Die unabhängige Genese dieser Systeme ist dafür verantwortlich, dass es trotz funktionaler Gemeinsamkeiten strukturelle Unterschiede zwischen ihnen gibt.

Drei in mancher Hinsicht verwandte Technologien sind Zahlensysteme, Kalender und Musiknotationen. Sie sind mit Schriften insofern vergleichbar, als dass sie eine wesentliche Erweiterung der kognitiven Leistungsfähigkeit ermöglichen. Bezüglich dieser erweiterten Leistungsfähigkeit sind diese Technologien im Laufe der Zeit verbessert worden, denn Technologien, die uns nicht helfen, Aufgaben besser oder überhaupt zu meistern, sind nutzlos. Die Qualität von Technologien sollte sich dementsprechend an ihrer relativen Nützlichkeit bemessen.

1. Zahlen

Prototypen einer Technologie des Geistes sind Zahlensysteme. Natürliche Objekte, Artefakte und Personen zu zählen und darauf aufbauend Mengen miteinander in Beziehung zu setzen, gehört zu den fundamentalen geistigen Tätigkeiten, um die Welt zu verstehen. Alle Kulturen haben Zahlensysteme entwickelt, die sich freilich nach ihrer Beschaffenheit und den über das Kopfrechnen hinausgehenden Rechenoperationen, die sie ermöglichen, sehr voneinander unterscheiden. Sie reichen von einfachen Additionssystemen, in denen jeder Wert mit Symbolen — z.B. Strichen — dargestellt wird, deren Anzahl der gezählten Objekte entspricht, bis zu komplexen Stellenwertsystemen, in denen sich der Wert eines Symbols nach einer konventionell festgelegten Reihenfolge bestimmt. Denken bedeutet nach Immanuel Kant, einzelne Wahrnehmungen allgemeinen Begriffen zuzuordnen und so zu Urteilen zu kommen. Dass ein Baum, ein Schaf und ein Haus gleichermaßen die Zahl Eins verkörpern und ein Zählschritt zwei Bäume, Schafe etc. diesem allgemeinen Schema zuordnet, ist eine Abstraktion, also eine Erkenntnisleistung.

Berechnungen sind keine direkten Abbildungen gegebener numerischer Verhältnisse, sondern Produkt der Abstraktions- und Verallgemeinerungsmöglichkeiten, die ein Zahlensystem bietet. Eine etwas kompliziertere Abstraktion als die Zahlen von Objekten der Anschauung sind z.B. negative Zahlen. Wenn du mir drei Scheffel Getreide zurückgibst, ich dir aber vier geliehen hatte, ist das einer zu wenig. Wie unterscheidet sich dieser eine von einem der drei zurückgegebenen Scheffel? Das ist eine Frage der symbolischen Darstellung, ohne deren Beantwortung es

nicht möglich ist, in komplexeren Berechnungen negative gegen positive Zahlen aufzurechnen. Im heute für den Alltag gebräuchlichen Zahlensystem geschieht das durch ein einfaches Vorzeichen: -1 , -2 , -3 gegenüber 1 , 2 , 3 . Vielen historischen Zahlensystemen fehlt ein entsprechendes symbolisches Mittel, was sie für komplexe Berechnungen, die negative Zahlen beinhalten, ungeeignet macht.

Die größte und für die Entwicklung der Mathematik folgenreichste Abstraktion ist die Null (Seife, 2000). Im Laufe der Geschichte wurde ihr dreimal ein eigener Platz in Zahlensystemen gegeben, in Babylon, in Mittelamerika von den Mayas und in Indien (Cajori, 1929). Die Idee, dass mit Nichts gerechnet werden muss und es deshalb ein eigenes Zeichen haben muss, wurde explizit von Brahmagupta begründet, der im siebten Jahrhundert u.Z. Rechenoperationen mit Zahlzeichen für 1 bis 9 und der Null formalisierte. Abu Dscha'far Muḥammad ibn Mūsā al-Khwārizmī, latinisiert: *Algorismi* (< Algorithmus) führte 825 die indische Null in die Welt arabischer Gelehrsamkeit ein, von wo sie langsam nach Byzanz, Spanien und den Rest von Europa kam. Als Leonardo Fibonacci diese geniale Erfindung im zwölften Jahrhundert u.Z. nach Italien brachte, stieß sie zunächst auf Skepsis. Denn in dem so entstandenen dezimalen Stellenwertsystem hatte die Null die überraschende Eigenschaft, dass sie alleinstehend keinen Wert darstellte, aber in einer Zahlenfolge den Wert der ihr unmittelbar vorausgehenden Ziffer verzehnfachte.

Mathematiker erkannten zwar die Überlegenheit des dezimalen Stellenwertsystems schnell, aber es sollte noch Jahrhunderte dauern, bis es sich allgemein durchgesetzt hatte. Für das Aufzählen, das nicht unbedingt ein Stellenwertsystem benötigt, blieben die römischen Zahlen noch länger in Gebrauch.

Ähnlich verhielt es sich in China, wo das indisch-arabische Zahlensystem im dreizehnten Jahrhundert von den muslimischen Hui eingeführt wurde. Auch hier hielt man weiter an dem gewohnten Hybridsystem fest, das keine Null kennt.

Selbst für relativ einfache Berechnungen sind die römischen ebenso wie die chinesischen Zahlen unpraktisch, wie die Darstellung einer beliebigen Zahl, z.B. 44 verdeutlichen mag.

- Striche. Die Gesamtheit entspricht der Summe:
<II> (44 Striche).
- Chinesisch. Die Ziffernfolge <四十四> entspricht der Folge $4\ 10\ 4$ und ist als $(4 \times 10) + 4$ zu interpretieren.
- Römisch. Die Ziffernfolge <XLIV> entspricht der Folge $10\ 50\ 1\ 5$ und ist als $(50 - 10) + (5 - 1)$ zu interpretieren.
- Indisch-arabisch. Die Ziffernfolge <44> ist als $40 + 4$ zu interpretieren.

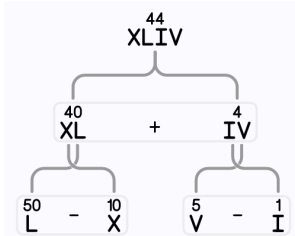


ABBILDUNG 1. Die römische Zahl XLIV (44) aufgeschlüsselt nach Position und Wert

Eine einfache Multiplikation wie $4.799 \times 3 = 14.397$ wird mit chinesischen und römischen Zahlen zu einer Herausforderung:

- Chinesisch. 四千七百九十九 · 三 = 一万四千三百九十七
- Römisch. MMMDCCXCIX · III = $\overline{\text{X}}\overline{\text{M}}\overline{\text{V}}\overline{\text{C}}\overline{\text{C}}\overline{\text{C}}\overline{\text{X}}\overline{\text{C}}\overline{\text{V}}\overline{\text{I}}$

Die praktischen Vorteile des Rechnens mit der Null und dem dezimalen Stellensystem erwiesen sich schließlich als unabweisbar und ließen die arabischen Zahlen zum universellen Standard werden. Dieser fragmentarische Hinweis auf die Geschichte der Zahlensysteme muss an dieser Stelle genügen, um die kognitiven Konsequenzen einer Technologie des Geistes und das komplexe Verhältnis zwischen scheinbar gegebenen, also durch Anschauung überprüfbareren Daten und ihrer Darstellung zu illustrieren.

2. Kalender

Eine weitere intellektuelle Technik, die diesen Zusammenhang verkörpert und auf Zahlensystemen aufbaut, ist der Kalender. Wie wir ihn heute kennen, ist er das Ergebnis jahrhundertelangen Ringens mit der Zeit, um ihren Verlauf in Intervalle zu fassen und auf regelmäßige und für den Menschen relevante Zyklen abzubilden, insbesondere das Jahr (Duncan, 1998). Diese Geschichte lässt sehr deutlich erkennen, dass das Instrument den Gegenstand, den es messen soll, als Bewusstseinsstatuse erschafft. Wie schwierig es ist, Zeit zu messen, offenbart sich darin, dass wir über die zeitliche *Erstreckung* kaum reden können, ohne uns räumlicher Metaphern zu bedienen. Um zu messen, brauchen wir Maßeinheiten, ein Lichtjahr zum Beispiel. Ist das ein räumliches oder ein zeitliches Maß? Diese Frage nach Metrik und Topologie der Zeit soll uns hier nicht weiter beschäftigen. Es genügt die Feststellung, dass wir sie gar nicht stellen könnten, wenn wir nicht wüssten, was ein Jahr ist oder zumindest dächten, dass wir es wüssten.

Die Verbegrifflichung der Zeit vollzog sich im Spannungsverhältnis zwischen Almanach und Chronik, Astrologie und Astronomie. Die

Verbindung der mechanischen Uhr mit astronomischen Beobachtungsscheiben war dabei eine wichtige technische Neuerung, die auf das Verhältnis von Vergänglichkeit und Wiederholung und den Zyklus der Gestirne verwies. Die Berechnung ihrer Laufbahnen wurde immer genauer, was Verbesserungen der in verschiedenen Teilen der Welt entstandenen Kalender nach sich zog. Sonne, Mond und Sterne boten sich als Bezugspunkte an, um den kontinuierlichen Lauf der Zeit in diskrete Sonnenjahre, Mondjahre oder siderische Jahre einzuteilen. Der Kalender steht so direkt mit unserer Kosmologie in Verbindung, mit dem was wir über die Kräfte, die die Himmelskörper bewegen, wissen und vermuten.

Die genaue Länge des Jahres ist der Rahmen, in den natürliche und soziale Rhythmen eingepasst sind – Jahreszeiten, Monate, Wochen, Arbeits-, Bet- und Feiertage, Stunden, Minuten, Sekunden – die den Kalender zu einem Herrschaftsinstrument machen, über das der Herr der Zeit verfügt. Alternative Kalender bedrohen seine Autorität und werden deshalb nicht leicht akzeptiert. Schon 1596 hatte der Jesuit Matteo Ricci in Beijing den Zeitpunkt einer Sonnenfinsternis viel genauer vorausberechnet als chinesische Astrologen/Astronomen, die einen luni-solaren Kalender verwendeten (Needham, 1981, S. 68). 1665 wiederholte Riccis Kollege Johann Adam Schall diesen Beweis der Überlegenheit der europäischen Astronomie und Zeitmessung (Udias, 1994). Es dauerte gleichwohl noch lange, bis man in China die Zeit ad anno Domini maß, sollte der dominus doch nicht der Herr der Chinesen sein. Letzten Endes jedoch setzte sich der Gregorianische Kalender durch, wie er auch den Julianischen Kalender (mit dem Julius Caesar 45 v.u.Z. den alten römischen Kalender ersetzt hatte) und Kalender in anderen Teilen der Welt verdrängte. Er war besser, genauer. So reduzierte er die elfminütige Abweichung vom tropischen Jahr, die Caesars Kalender aufwies, auf 26 Sekunden im Jahr; ein unbezweifelbarer Fortschritt. Allein, nüchterne Funktionalität setzt sich nicht immer sofort durch, denn die Astronomie hat die Astrologie nie ganz aus dem Feld geschlagen, und irrationale Abneigungen stehen rationalen Entscheidungen oft im Weg. Im protestantischen Schweden wurde der 1582 verkündete Kalender des Katholiken Gregor erst 1753 übernommen und in Russland sowie anderen Ländern der orthodoxen Welt dauerte es bis ins zwanzigste Jahrhundert. Auch die Chinesen erkannten die Überlegenheit von Matteo Riccis Berechnung, hielten aber dennoch an ihrem lunisolaren Kalender fest, bis Mao Zedong bei der Gründung der Volksrepublik China erklärte, dass Jahre fortan wie im Westen gezählt würden.

3. Noten

Das dritte, hier nur in aller Kürze zu erwähnende Beispiel einer Abbildungstechnik, die auf Repräsentation und Wahrnehmung ihres Gegenstands einwirkt, sind Notensysteme. Auch hier finden wir die frühesten

Versuche in den alten Hochkulturen, bei den Sumerern, bei den Griechen und bei den Mayas, zum Beispiel (Schneider, 1987). Die Musiknotation, die heute am gebräuchlichsten ist, ist das Ergebnis jahrtausendelanger Bemühungen, das Klangkontinuum zu gliedern und visuell darzustellen (Gould, 2011). Dabei wurden sukzessive mehr Parameter bzw. Dimensionen von Schallereignissen miteinbezogen: Tonhöhe, Lautstärke, Dauer, Klangfarbe und Rhythmus. Wiedergabe und Weitergabe geschätzter Hymnen waren anfänglich vermutlich der Zweck, Musiknotationen zu entwickeln. Analyse und Harmonielehre kamen später dazu und schließlich die Komposition. Das menschliche Gehirn ist fähig, lange Melodien nach Gehör im Gedächtnis zu behalten und wiederzuerkennen; komplexe Kompositionen werden jedoch erst durch geeignete Notationen ermöglicht. *Die Kunst der Fuge* von Bach, die *Eroica* von Beethoven, die *Bilder einer Ausstellung* von Mussorgski wären ohne chromatische Notation nie entstanden. Das syntaktische Fünf-Linien-System mit Notenschlüsseln fungiert also gleichermaßen als Abbild und als Vorbild, denn es erlaubt nicht nur, Melodien festzuhalten, sondern macht bestimmte komplexe Tonkonfigurationen erst möglich.

Die Kulturen der Menschheit haben im Laufe der Geschichte viele verschiedene Systeme zur visuellen Darstellung von Musik hervorgebracht, und die Entwicklung ist nicht abgeschlossen. Komponisten fahren fort, sie zu verfeinern und experimentieren weiter, denn sie wissen, dass ihr Handwerkszeug profunden Einfluss auf ihr Produkt hat.

4. Schrift

Die für die Menschheitsgeschichte wichtigste Technologie des Geistes, die Schrift, teilt manche Eigenschaften mit den erwähnten drei Beispielen. Zahlen schaffen Ordnung im Universum; Kalender definieren Perioden, die den kontinuierlichen Lauf der Zeit gliedern; und Noten machen Harmonie sichtbar, indem sie flüchtige Klänge symbolisch festhalten. Auch Schrift artikuliert ein in direkter Sinneswahrnehmung fließendes Ereignis, das ohne sie im Augenblick seiner Realisierung vergeht. Auf die produktive Kraft von Schriftsystemen haben aus verschiedenen Blickwinkeln neben vielen anderen Ong (1982), Olson (1994), Harris (2000), Coulmas (2003) und Gnanadesikan (2009) hingewiesen. Schrift macht Formen des Sprachgebrauchs möglich, die es in einer mündlichen Kultur nicht gibt, z.B. Register, Listen, Briefe, heilige Bücher, Romane, Promessen, Steuererklärungsformulare, E-Mail-Adressen, Tweets, um nur einige zu nennen. Poeten können auch ohne Schrift reimen, Verse schmieden und deklamieren, aber „mündliche Literatur“ ist ein Oxymoron. *Genji monogatari*, *Ulysses*, *À la recherche du temps perdu* und *Die Buddenbrooks* haben in einer mündlichen Kultur keine Entsprechung.

Überall, wo voll ausgebaute Schriftsysteme entstanden, entwickelten sich auch literarische und andere schriftsprachliche Genres. Die Frage, die hier zumindest gestellt, wenn auch nicht letztgültig beantwortet werden soll, ist, ob es diesbezüglich Qualitätsunterschiede zwischen Schriftsystemen gibt? Sind manche Schriftsysteme besser als andere, so, wie arabische Zahlen besser sind als römische und chinesische, und so, wie der Gregorianische Kalender besser ist als der Julianische? Wenn Schrift eine Technologie ist, die bestimmte geistige Tätigkeiten ermöglicht und ihrer Ausführung dient, ist das eine sinnvolle Frage, eine Frage, die wiederholt gestellt worden ist.

In ihrer Ausgabe vom 12.05.2016 überschrieb die durchaus renommierte Zeitschrift *Foreign Policy* einen Artikel: „Chinese is not a backward language“. Gemeint war nicht die Sprache, sondern die Schrift, und es wurde offenkundig für mitteilenswert gehalten, dass diese nicht unterentwickelt sei. Tatsächlich steht die Vermutung, der die Überschrift widerspricht, in einer ehrwürdigen Tradition, nach der die Geschichte der Schrift eine teleologische Entwicklung beschreibt, die mit dem griechischen bzw. lateinischen Alphabet ihren Höhepunkt erreicht. Diese Idee ist eine tragende Säule des Eurozentrismus, auf der lange der Dachfirst der geistigen Überlegenheit der weißen Rasse bzw. der europäischen Kolonialmächte ruhte. Ein prominentes Beispiel ist Hegels diesbezügliche Einlassung. Im Zusammenhang mit Überlegungen zu der geistigen Bedeutung von Zeichen vergleicht er die Güte von Hieroglyphen- und Buchstabenschrift und konstatiert:

Die Buchstabenschrift ist an und für sich die intelligentere; in ihr ist das Wort, die der Intelligenz eigentümliche würdigste Art der Äußerung ihrer Vorstellungen, zum Bewusstsein gebracht, zum Gegenstande der Reflexion gemacht. (Hegel, 1830, III §459)

Einer der Hauptgründe, die Hegel für sein Urteil anführt, ist, dass abstrakte Begriffe in einer Hieroglyphenschrift nicht darstellbar sind und somit auch nicht „zum Gegenstand der Reflexion gemacht“ werden. Namen, Eigennamen ebenso wie Namen solcher Begriffe, sind nach Hegel (ibid.) „für sich sinnlose Äußerlichkeiten“. Die Buchstabenschrift suggeriert nichts anderes, während die Bildhaftigkeit der Hieroglyphen die Reflexion auf den abgebildeten Gegenstand lenkt statt auf den abstrakten Begriff.

Mit der Bewertung der verschiedenen Instrumente lässt Hegel es nicht bewenden, denn nach seiner durch Wilhelm von Humboldt beeinflussten Auffassung besteht zwischen ihnen und der geistigen Konstitution ihrer Benutzer ein Zusammenhang. Bezüglich der chinesischen Schrift stellt er fest: „Nur dem Statarischen der chinesischen Geistesbildung ist die hieroglyphische Schriftsprache dieses Volkes angemessen“ (ibid.). Was genau Hegel sich unter einer Hieroglyphenschrift vorstellte, wissen wir nicht genau, aber die Vermutung liegt nahe, dass er

ein Zeichensystem vor Augen hatte, dessen Elemente sich alle und ausschließlich kraft ihrer Bildhaftigkeit auf ihre Bezugsobjekte beziehen, denn er nimmt direkt auf Leibniz Projekt einer *Characteristica universalis* Bezug, das er als unrealisierbar ablehnt. Damit hatte Hegel zwar recht, er irrte aber in der Annahme, dass die chinesische Schrift oder andere „Hieroglyphenschriften“ wie etwa die ägyptische so funktionieren. Das kann man ihm nicht vorwerfen, aber dass er auf der Grundlage mangelhafter Kenntnisse weitreichende Schlüsse über Schriftsysteme, ihre Benutzer und die geistigen Dispositionen von Völkern zog, enthüllt den ideologischen Charakter dieser Übung.

Wie ideologisch ist die auch heute verbreitete Hochschätzung „des Alphabets“? Ist es möglich, zu einer unideologischen Qualitätsbewertung von Schriftsystemen zu kommen? Eine entsprechende Frage bezüglich Sprachen zu stellen, hieße, ein Tabu der Linguistik zu verletzen, da die Idee von der „natürlichen Sprache“ tief verwurzelt ist und zu dem Grundsatz zwingt, dass alle Sprachen gleich gut sind. Schriftsysteme sind hingegen ohne Zweifel keine natürlichen Objekte, sondern Artefakte, Techniken, die man auf ihre Qualität beurteilen könnte. Schwierig wird das dadurch, dass Schriften mit Sprachen verbunden sind und deshalb nicht *sui generis* beurteilt werden können, obwohl das manchmal behauptet wird.

5. Bewertungskriterien

Die Geschichte der Schrift ist u.a. auch die Geschichte der Schriftreformen, die einen Wechsel des Schriftsystems beinhalten können – z.B. die Ersetzung der chinesischen durch die Lateinschrift in Vietnam – oder, weniger dramatisch, Anpassungen der Orthographie betreffen wie z.B. die Ersetzung der polytonischen durch die monotonische amtliche Schreibweise des Griechischen 1982. Wenn um solche Reformen gestritten wird (und gestritten wird dabei unweigerlich), führen Befürworter und Gegner stets unwiderlegbar Qualitätsargumente ins Feld, deren theoretische Untermauerung ihnen Experten liefern.

Sieben in diesem Zusammenhang wiederholt vorgeschlagene Bewertungskriterien sind die folgenden:

1. visuelle Verarbeitbarkeit
2. kognitive Verarbeitbarkeit
3. sprachliche Angepasstheit
4. intersprachliche Übertragbarkeit
5. Erlernbarkeit (Schreiben und Lesen)
6. ästhetische Wirkung
7. Kontinuität der Tradition

Die ersten fünf Kriterien sind utilitaristischer, die letzten beiden kultureller Art.

1. Schriften sind visuelle Notationssysteme, die aus einem Repertoire von Grundzeichen bestehen. (Ich ignoriere hier Braille und andere taktile Schriftsysteme, für die aber Ähnliches gilt.) Das Repertoire muss erschöpfend sein, also dazu geeignet, alles auszudrücken, was ausgedrückt werden soll; und jedes der Zeichen muss sich von jedem anderen unterscheiden. Zu kleine graphische Unterschiede bergen das Risiko, Unverständnis oder Missverständnisse seitens der Empfänger zu verursachen. Graphische Überdifferenzierung vergrößert andererseits den Aufwand der Zeichenproduktion unnötig. Schriften müssen ein gewisses Maß an Redundanz aufweisen, um eine Balance zu finden und beide Risiken zu minimieren. Unterscheiden sich Schriften bezüglich des Maßes der ihnen inhärenten Redundanz? Die Forschungsergebnisse, die hierzu vorliegen, deuten eher auf Gemeinsamkeiten hin als auf Differenzen. So haben z.B. Changizi und Shinzuke (2005) mehr als hundert Schriften verglichen und nur geringe Redundanzunterschiede festgestellt, was sie damit begründen, dass Schriftsysteme „dem Selektionsdruck ausgesetzt seien, Zeichen zu verwenden, die für das visuelle System leicht erkennbar sind“.
2. Redundanz ist auch ein Kriterium der kognitiven Verarbeitbarkeit, aber nicht das einzige. System spezifische Eigenschaften kommen hinzu. Verkompliziert wird die Bewertung dieser Variablen dadurch, dass sie sich auf Rezeption und Produktion auswirkt. Ein rein phonetisches System ist einfach fürs Schreiben — Homophone werden gleich geschrieben — aber für das Leseverständnis ist ein phonosemantisches System einfacher (Leib/Laib, bar/Bar, etc.). Praktisch alle Schriftsysteme sind gemischte Systeme (was Hegel nicht zur Kenntnis nahm) und unterscheiden sich im Verhältnis laut- und bedeutungsbezogener Bestandteile nur graduell. Verlässliche Untersuchungen, die zeigen, dass eine Schrift bzw. ein Schriftsystem für die kognitive Verarbeitung einfacher ist als andere, fehlen, obwohl gut erforscht ist, dass bedeutungsorientierte und lautorientierte Schriftsysteme neuronal unterschiedlich gespeichert werden und ihre kognitive Verarbeitung durch Hirnverletzungen auf unterschiedliche Weise geschädigt werden kann (Leong und Tamaoka, 1998).
3. Wenn man in Aristotelischer Tradition davon ausgeht, dass Schriften Sprachen abbilden sollen — was keine zwingende Annahme ist — dann ist die Angepasstheit einer Schrift, d.h. die möglichst direkte Abbildung sprachlicher Einheiten ein Kriterium ihrer Güte. Semitische Schriften, die trikonsonantische Wurzeln semitischer Lexeme abbilden, und japanische Hiragana, die die Struktur japanischer Moren ziemlich getreu wiedergeben, sind in diesem Sinne gut angepasst. Aber was besagt dieses Kriterium für sich genommen? Es steht mit dem folgenden in Zusammenhang.

4. Man kann Dtsch hn Vkl schrbn, aber sehr praktisch ist das nicht. Konsonantenschriften sind für manche Sprachen geeigneter als für andere, also nicht so leicht transferierbar. Der Vorschlag, indoeuropäische Sprachen, sagen wir Sizilianisch, Bayrisch oder Westfriesisch, chinesisch zu verschriften, stieß wahrscheinlich unter Hinweis auf die gute Angepasstheit der chinesischen Schrift an die chinesische Sprache, aber schlechte Übertragbarkeit auf andere Sprachen auf Ablehnung. Als großer Vorteil des griechisch-lateinischen Alphabets wird demgegenüber häufig seine „universelle“ Übertragbarkeit gerühmt, weswegen sie „an und für sich die intelligentere ist“, wie Hegel sagte. Wie ist es darum bestellt? Dieser Nimbus beruht auf der Vorstellung, das Alphabet sei eine Art IPA und die Einheiten, die es abbildet seien Naturtatsachen. Das stellt jedoch die wahren Verhältnisse, wie Faber (1992) überzeugend nachgewiesen hat, auf den Kopf. Nicht das Alphabet ist eine Art IPA, sondern das IPA ist ein Alphabet, das die Spuren seiner Herkunft mit sich trägt. Es entstand im Kontext und für die Darstellung bestimmter Sprachen und ist deshalb z.B. bezüglich Vokalen unterdeterminiert, was bei vokalreichen Sprachen zur Verwendung von Diakritika zwingt (ä, ü, ö u.a.). Darüber hinaus lässt es einen ganzen Parameter der Lautsprache völlig unberücksichtigt, bedeutungstragende bzw. unterscheidende Töne, wie wir sie in Khoisan-Sprachen im südlichen Afrika und in sinotibetischen Sprachen finden. Vietnamesisch ist ein einschlägiges Beispiel. Töne sind Qualitäten von Silben. Da Silben als sprachliche Einheiten in der Alphabetschrift keine Existenz haben, werden Töne z.B. in lateinschriftlich transliterierten chinesischen Texten behelfsmäßig mit Diakritika auf Vokalbuchstaben wiedergegeben. Dieses Hilfsmittel fand faute de mieux Eingang ins vietnamesische Standardalphabet, was zur Folge hat, dass es 18 <a>, 12 <e>, 18 <o>, 13 <u> und 10 <i/y> hat; extrem dysfunktional. Hier zeigt sich, dass das Lateinalphabet für diesen Sprachtyp sehr ungeeignet ist und sich bezüglich der Übertragbarkeit auf andere Sprachen nicht unbedingt vor anderen Schriften auszeichnet. Mit Sicherheit gilt das nicht für isomorphe Systeme wie die Kyrillica.

Die Schrift erweist sich als ein System, das Eigenschaften hat wie die anderen oben erwähnten Kulturtechniken auch: Indem es Kategorien abbildet, schafft es sie. Interessanterweise erahnte Hegel das in seinen Bemerkungen über die „an und für sich intelligentere Buchstabenschrift“; jedenfalls kann die folgende Bemerkung so gedeutet werden: „Die Ausbildung der Tonsprache hängt zugleich aufs genaueste mit der Gewohnheit der Buchstabenschrift zusammen, durch welche die Tonsprache allein die Bestimmtheit und Reinheit ihrer Artikulation gewinnt“ (Hegel, 1830, §459). Mit anderen Worten, wer nicht lesen und schreiben kann, spricht auch nicht klar artikuliert. Anders als für Ferdinand de Saussure und Linguisten, die ihm folgten, war für Hegel *la tyrannie de la lettre*, will sagen, der Einfluss der Schrift auf

die Sprache, kein Problem (Coulmas, 2018). Im Gegenteil, einen solchen Einfluss auf die Ausbildung der Tonsprache geltend machen zu können, war einer der Gründe dafür, dass er die Buchstabenschrift für überlegen hielt. Im hier gegebenen Zusammenhang ist jedoch wichtiger, dass diese normative Kraft des Lateinalphabets, wie das vietnamesische Beispiel zeigt, nicht bei allen Sprachen gleichermaßen zur Geltung kommen kann.

5. „So simpel wie das ABC“ – diese Redensart überzeugt im Falle des vietnamesischen Alphabets nicht so recht, und bei genauerer Betrachtung ist sie im allgemeinen sehr irreführend, denn nicht die oft hervorgehobene Ökonomie des Inventars der Grundzeichen bestimmt die relative Komplexität einer Schrift, sondern die Regeln ihrer Kombination. Wenn wir die 26 Buchstaben des Lateinalphabets mit den rund 2.500 chinesischen Zeichen für den alltäglichen Gebrauch vergleichen, ist klar, das ABC ist leichter zu erlernen als die chinesischen Zeichen. Ganz so einfach sind die Verhältnisse allerdings nicht, da sich Alphabetschriften sehr stark hinsichtlich der Komplexität ihrer Orthographien unterscheiden und somit bezüglich ihrer Erlernbarkeit. Italienisch und Finnisch sind einfach, Englisch und Französisch schwierig. Italienische Kinder lernen schneller lesen als englische, da die englische Orthographie sehr komplex ist (Thorstadt, 1991). Um Englisch flüssig lesen und schreiben zu können, muss man ca. 1.800 Phonem-Graphem-Kombinationen gelernt haben, eine Größenordnung, die mit den 2.500 chinesischen und 2.136 japanischen Zeichen für den normalen Gebrauch vergleichbar ist. Nach einer teleologischen Vorstellung von der Geschichte der Schrift geht die Entwicklung mit der sukzessiven Verringerung des Zeicheninventars und der Verkleinerung der abgebildeten sprachlichen Einheiten einher, was als Ergebnis des von Changizi und Shinsuke (2005) erwähnten Selektionsdrucks gedeutet werden könnte. Dass die Menge der relevanten Einheiten, etwa der Phonem-Graphem-Kombinationen, im Laufe der Zeit größer wird, ist dabei nicht vorgesehen. Tatsächlich müssen wir aber feststellen, dass die asynchrone Entwicklung von Sprache und Schrift bei lautbezogenen Schriften eben eine solche Vergrößerung nach sich zieht. Gleichzeitig stellen wir auch fest, dass die Menge der chinesischen Zeichen zwar prinzipiell offen ist, die des alltäglichen Gebrauchs in China und in Japan aber über die Jahrhunderte mehr oder weniger konstant geblieben ist. Die Tatsache, dass Japan und China den Vorsprung des Westens in Wissenschaft und Technik aufgeholt, ja, den Westen auf manchen Feldern überholt haben, diskreditiert die Glaubwürdigkeit des von Hegel angenommenen Zusammenhangs zwischen „dem Statarischen der chinesischen Geistesbildung“ und der chinesischen Schrift und wirft die Frage auf, ob die Ökonomie des Inventars der Grundzeichen einer Schrift und ihre

sprachliche Bezugseinheit überhaupt eine relevant Variable für die Qualitätsbewertung von Schriften ist.

6. Unter der Prämisse des Ranking & Rating, das im neoliberalen Zeitalter in alle Lebensbereiche eingedrungen ist, gerät leicht in Vergessenheit, dass dieses utilitaristische Denken eine rezente Erscheinung ist. Es favorisiert Funktionalität, Faulheit (das „Prinzip des geringsten Aufwands“) und rationale Entscheidung, lässt aber darunter nicht fassbare Werte außeracht. Wie wahrscheinlich wäre es, dass Lafcadio Hearn's Beurteilung japanischer Schriftlichkeit von 1890 heute ernstgenommen würde?

Und schließlich ... wird dich wie eine Offenbarung das Bewusstsein überkommen, dass der erstaunliche malerische Reiz dieser Straßen einfach nur in der Fülle der japanischen und chinesischen Schriftzeichen liegt. ... Vielleicht dass du dir dann für einen Augenblick die Wirkung vergegenwärtigst, die es hätte, wenn an Stelle dieser magischen Zeichen das lateinische Alphabet gesetzt würde — und die bloße Idee wird dir einen heftigen Ruck geben, und du wirst gleich mir ein Feind der „Romaji-Kwai“ werden, jener für den hässlichen utilitaristischen Zweck gegründeten Gesellschaft zur Einführung lateinischer Buchstaben in die japanische Schrift.

(Hearn, 1823, S. 19)

Wer an das Prinzip des geringsten Aufwands glaubt, an Selektionsdruck und daran, dass ein kleines Zeicheninventar ein Vorteil ist, kann über Hearn's Schwärmerei von den „magischen Zeichen“ nur lächeln. Es ist wahrscheinlich, dass er Hegel's Hochpreisung der Buchstabenschrift nicht kannte, und es ist ebenfalls wahrscheinlich, dass Hearn's Verständnis von der Funktionsweise der chinesischen Schrift nicht richtiger war als das Hegel's. Interessant ist freilich, dass Hearn eben die Eigenschaften des Alphabets schmätzt, die Hegel lobt, nämlich dass es aus „unbelebten, trockenen Symbolen von Stimmlauten“ (ibid.) besteht. Wenn die Menschen sich eine Schrift aussuchen könnten und diesbezüglich im Sinne der Theorie der rationalen Entscheidung (Kahneman, 2003) die Wahl zwischen Hegel's und Hearn's Position hätten, wo läge die Mehrheit? Das wissen wir nicht, da es sich um eine hypothetische Frage handelt. Es ist aber sicher, dass sich viele Menschen bezüglich Ästhetik auf Hearn's Seite schlagen und seine Meinung zu „dem hässlichen utilitaristischen Zweck der Gesellschaft zur Einführung lateinischer Buchstaben“ teilen würden. Die Tatsache, dass sich Japan trotz zweimaligen äußerst starken Drucks — in der Epoche der Modernisierung im ausgehenden neunzehnten Jahrhundert und nach dem Zweiten Weltkrieg seitens Amerikas — nicht zur Übernahme des Lateinalphabets entschlossen hat, ist nur ein Indiz. In vielen Schriftkulturen werden Schriften nach ihrer ästhetischen Wirkung beurteilt, ebenso wie Sprachen, wie irrational das auch sein mag. Die Ästhetik ist ein Faktor, der die Wirksamkeit des Prinzips des geringsten Aufwands einschränkt, aber nicht der einzige.

7. Was Schriftlichkeit von Mündlichkeit unterscheidet, ist die Bewahrung und Weitergabe von Information unabhängig von Ort und Zeit. Es ist deshalb nicht überraschend, dass sich gegen Schriftreformen stets Widerstand erhebt, denn sie sind eine potentielle Bedrohung dieser Kernfunktion. Wer kann heute noch Osmanlıca lesen, die bis vor 100 Jahren konventionelle Schreibweise des Türkischen? Die Ersetzung der arabischen durch die lateinische Schrift in den 1920er Jahren war ein Traditionsbruch, wie er selten vorkommt, eben weil Dauerhaftigkeit ein Wesensmerkmal der Schrift ist und Tradition von vielen als Wert an sich begriffen wird. Deshalb ist es so unwahrscheinlich, dass die Chinesen die chinesische Schrift aufgeben; deshalb wird die insgesamt geringfügige Schriftzeichenreform der VR China in den 1960er Jahren in Taiwan als Sakrileg betrachtet; deshalb denkt in Griechenland niemand daran, die Lateinschrift zu übernehmen (obwohl sich Griechisch oder Λατινοελληνικά kurzfristig im Cyberraum zu etablieren schien). Wie Ästhetik spielt Traditionsbewahrung im Ensemble der Kriterien für die Bewertung von Schriften und Schriftsystemen eine unabhängige Rolle. Zuverlässige Forschungsergebnisse dazu, inwieweit verschiedene Schriften der Traditionsbewahrung in dem Sinne dienlich sind, dass sie die Zugänglichkeit zu ältere Sprachstufen erhalten, sind nicht bekannt. Zu diesem Zweck müsste z. B. ein Maß gefunden werden, um zu beurteilen, ob Leser der entsprechenden Sprachen — z.B. Deutsch und Chinesisch — heutzutage das Vaterunser oder das Gedicht von Li Bai, beide aus dem achten Jahrhundert u.Z., leichter verstehen können.

<p>Fater unsêr, thû pist in himile, uuîhi namun dînan, quueme rîhhi dîn, uuerde uuillo dîin, sô in himile sôsa in erdu. Prooth unsêr emezzihic kip uns hiutu, oblâz uns sculdi unsêro, sô uuir oblâzêm uns sculdîkêm, enti ni unsih firleiti in khorunka, ûzzer lôsi.</p>	<p>对酒不觉暝， 落花盈我衣。 醉起步溪月， 鸟还人亦稀。</p>
---	--

Nur mit quantitativen Tests könnte ein empirisch belastbares Maß entwickelt werden. Angesichts der verschiedenartigen Variablen, die dabei zu berücksichtigen wären — Textart und -länge, Sprachtyp, Bildungstradition, Literalisierungs- und Bildungsniveau der Gesamtbevölkerung, etc. — ist es unwahrscheinlich, dass ein Vergleich, der den Einfluss des Schriftsystems auf die Zugänglichkeit des Schrifttums älterer Sprachstufen objektiv ermisst, überhaupt möglich ist.

6. Viel Lärm um nichts

Welche Schlussfolgerungen lassen sich aus der Synopse der angesprochenen sieben Kriterien zur Qualitätsbewertung von Schriftsystemen und ihrem diesbezüglichen Vergleich mit anderen intellektuellen Technologien ziehen? Obwohl für dekorative Zwecke nach wie vor verschiedene

Zahlensysteme, Kalender und Musiknoten in Gebrauch sind, hat sich die Welt weitgehend darauf geeinigt, dass arabische Zahlen, der Gregorianische Kalender und die chromatische Musiknotation anderen Systemen funktional überlegen sind und dass die Funktionalität dieser Technologien für ihre Verwendung entscheidend ist. Bei Schriftsystemen ist das nicht der Fall. Dass nach mehr als fünf Jahrtausenden Schriftgeschichte noch immer mehrere hundert Schriften in Gebrauch sind, zeugt davon. Die Gründe dafür sind ebenfalls vielfältig.

Wie bei anderen Technologien auch, macht sich Pfadabhängigkeit bemerkbar, d.h. die Neigung, in der ausgefahrenen Spur zu bleiben, auch wenn ein anderer Weg näher oder bequemer ist. Die Qwerty-Tastatur ist ein einschlägiges Beispiel. Für die mechanische Schreibmaschine erfunden, um das Verklemmen von Typenhebeln zu verhindern, wird sie auf PC-Tastaturen weiterverwendet.

Ein weiterer Faktor ist, dass Schrift praktisch überall mit Sprache identifiziert wird. Jeder weiß, dass alle Sprachgemeinschaften sprachen, bevor sie schrieben und alle Kinder sprechen lernen, bevor sie schreiben lernen. Dennoch verleiht für die meisten Menschen nur die schriftliche Norm den Status einer „richtigen“ Sprache. Von hier ist es ein kurzer Schritt zu der Überzeugung, dass man Japanisch „nicht wirklich“ mit lateinischen Lettern schreiben kann oder dass das <ß> ein Element der deutschen Sprache ist. Hier zeigen sich die Spätwirkungen der Mystifizierung alter Schriften – der göttliche Ursprung der hebräischen, arabischen und Brahmi-Schrift etc. – die sich gegen die ungehemmte Kraft des Utilitarismus in der Schriftentwicklung auswirken. Wie der Kalender im Namen des Herrn ist auch die Schrift bzw. die Orthographie ein Herrschaftsinstrument. Wie die Geschichte lehrt, braucht die Autorität, die damit ausgeübt wird, nicht unbedingt rational begründet zu sein.

Wenn wir die Schriften der Welt nach den obigen sieben Kriterien und vielleicht noch einigen anderen miteinander vergleichen, müssen wir außerdem zu dem Schluss kommen, dass die bessere Schrift, wenn es sie denn geben sollte, keinen Wettbewerbsvorteil darstellt. Entscheidend ist, dass die Technologie der Schrift verfügbar ist und nicht, welche Form sie im Einzelnen hat. Das japanische Schriftsystem wird häufig als das komplizierteste der Welt beschrieben. Das Lateinalphabet ist in seiner finnischen Orthographie äußerst leicht zu erlernen. Die koreanische Hangul-Schrift ist in systematischer Sicht viel raffinierter als die Lateinschrift, gleichviel in welcher Ausprägung, da sie mehr sprachliche Strukturebenen erkennen lässt und zudem ein hohes Maß an intersprachlicher Übertragbarkeit aufweist. Die chinesische Schrift ist in dem Sinne ökonomisch, als dass chinesische Texte stets kürzer sind, als Übersetzungen in Sprachen mit anderen Schriftsystemen.

Die Aufzählung und Abwägung der Vorzüge einzelner Schriftsysteme ließe sich leicht fortsetzen und durch eine gegenteilige Liste von Nachteilen einzelner Schriften ergänzen, angefangen z.B. mit der missratenen

Rechtschreibung des Englischen. Diese unnötig komplexe Orthographie hat freilich das Englische nicht daran gehindert, zur meistgelernten Fremdsprache der Welt zu werden, noch hat die geniale Hangul-Schrift in den letzten 500 Jahren irgendeine Sprachgemeinschaft dazu bewegen können, sie für ihre Sprache zu übernehmen.

Schriften sind wie Kalender und Zahlensysteme Techniken nicht nur der Abbildung, sondern der Darstellung. Es sind Projektionen, die ihren unsichtbaren Gegenstand sichtbar machen und ihm für diesen Zweck Kategorien auferlegen. Für ihre Benutzer stellen sie ihre Sprache dar und sind deshalb nicht nur austauschbares Medium, sondern die Verkörperung der Sprache. Das hat zur Konsequenz, dass utilitaristische Kriterien für die Entwicklung und Verbesserung von Schriften nie allein ausschlaggebend sind. Wäre die Buchstabenschrift tatsächlich an und für sich die intelligenterere, und wenn Intelligenz ein echter Wettbewerbsvorteil wäre, müsste sich diese Schrift durchgesetzt haben. Daraus, dass das einstweilen nicht der Fall ist, kann man schließen, dass alle Versuche die Technik zu verbessern, alle Schrift- und Orthographiereformen, ziemlich unnützlich sind: viel Lärm um nichts.

Literatur

- Cajori, Florian (1929). *A History of Mathematical Notation*. Bd. 1. Mineola: Dover Publications.
- Changizi, Mark A. und Shimojo Shinsuke (2005). “Character Complexity and Redundancy in Writing Systems over Human History”. In: *Proceedings Biological Sciences* 272, S. 267–275.
- Coulmas, Florian (2003). *Writing Systems. An Introduction to their Linguistic Analysis*. Cambridge: Cambridge University Press.
- (2018). “Revisiting the ‘Tyranny of Writing’”. In: *The Tyranny of Writing. Ideologies of the Written Word*. Hrsg. von Costanze Weth und Kasper Juffermans. London: Bloomsbury, S. 19–29.
- Duncan, David E. (1998). *Calendar. Humanity’s Struggle to Determine a True and Accurate Year*. New York: Avon Books.
- Faber, Alice (1992). “Phonemic segmentation as epiphenomenon: evidence from the history of alphabetic writing”. In: *The Linguistics of Literacy*. Hrsg. von P. Downing, S.D. Lima und M. Noonan. Amsterdam: Benjamins, S. 111–134.
- Gnandesikan, Amalia E (2009). *The Writing Revolution*. Oxford: Wiley-Blackwell.
- Goody, Jack (1977). *The Domestication of the Savage Mind*. Cambridge: Cambridge University Press.
- Gould, Elaine (2011). *Behind Bars—The Definitive Guide to Music Notation*. London: Faber Music.
- Harris, Roy (2000). *Rethinking Writing*. London: Athlone.

- Hearn, Lafcadio (1823). *Das Japanbuch. Eine Auswahl aus den Werken von Lafcadio Hearn*. Frankfurt a. M.: Literarische Anstalt Rütten & Loening.
- Hegel, Georg Wilhelm Friedrich (1830). *Enzyklopädie der Philosophischen Wissenschaften*. Heidelberg: Oßwald.
- Ifrah, Georges (1987). *Universalgeschichte der Zahlen*. Frankfurt a. M.: Campus-Verlag.
- Kahneman, Daniel (2003). "Maps of Bounded Rationality: Psychology for Behavioral Economics". In: *The American Economic Review* 93.5, S. 1449–1475.
- Leong, Che Kan und Katsuo Tamaoka (1998). "Cognitive Processing of Chinese Characters, Words, Sentences and Japanese Kanji and Kana: An Introduction". In: *Reading and Writing* 10, S. 155–164.
- Needham, Joseph (1981). *The Shorter Science and Civilisation in China*. Bd. 2. Cambridge: Cambridge University Press.
- Olson, David (1994). *The World on Paper*. Cambridge: Cambridge University Press.
- Ong, Walter S. (1982). *Orality and Literacy: The Technologizing of the World*. London: Methuen.
- Richards, E. G. (1998). *Mapping Time: The Calendar and its History*. Oxford: Oxford University Press.
- Schneider, Albrecht (1987). "Sound, Sprache, Schrift: Transkription und Notation in der vergleichenden Musikwissenschaft und Musikethnologie". In: *Zeitschrift für Semiotik* 9.3–4, S. 317–343.
- Seife, Charles (2000). *Zero, Biography of a Dangerous Idea*. New York: Viking Adult.
- Thorstadt, G. (1991). "The Effect of Orthography on the Acquisition of Literacy Skills". In: *British Journal of Psychology* 82, S. 527–537.
- Udias, Agustín (1994). "Jesuit Astronomers in Beijing, 1601–1805". In: *Quarterly Journal of the Royal Astronomical Society* 35, S. 463–478.

Open and Closed Writing Systems. Some Reflections

Marc Wilhelm Küster

Abstract. Traditionally writing systems are mainly seen through the lens of how they represent language. This article explores an alternative classification that is built on their key internal characteristics. Seeing writing systems as sets of signs, it postulates just two main categories, namely those that are essentially open and those that are fundamentally closed. However, in various periods of their existence they can oscillate between these extremes.

After setting out this theory and its rationale, the article exemplifies this hypothesis using specimens from different times and writing systems. It studies how the medium—manuscripts, printing press, and digital media—have at least temporarily transformed writing systems, creating semi-closed or semi-open systems. The age of Unicode finally has brought us emojis, characters that have the potential to open up Europe’s quintessentially closed writing systems.

1. Characteristics of Writing Systems

Writing is Visible Speech¹. It exists primarily to persist fleeting words. Few, if anybody involved in the study of writing would deny this strong link between spoken and written language. In fact, starting from Taylor (1883, Vol. 2) and Gelb (1963) onward most authors classify writing systems into at least three categories, initially primarily into logosyllabaries, syllabaries, and alphabets, later into more sophisticated taxonomies.² And while Gelb (ibid.) insisted that writing systems are systems of signs independent from the phonemes of the spoken language,

Marc Wilhelm Küster  0000-0002-2600-0717
3b, rue de Wormeldange
L-7390 Blaschette
marc@budabe.eu

1. DeFrancis (1989).

2. Gelb would call these three categories word-syllabic, syllabic, and alphabetic writing and saw them as necessarily in historic progression (the classification itself goes back to the amazing Taylor (1883, Vol. 1), who spoke of “verbal signs,” “syllabic signs” and “alphabetic signs”). Terminology can vary, and for good reasons—cf. e.g., Daniels (1996, 4ff) for a fuller discussion and the reason to treat abjads and abugidas as fourth and fifth categories. Scholars may disagree about the correct classification

he also defines those three categories mainly with reference to the way they represent spoken language.

This is a fruitful and important view on writing systems, but one focusing our attention exclusively on the way writing encodes language. This article argues that while this remains a valid view, it is not the only angle from which writing should be viewed.

In this perspective writing is much more than just Visible *Speech*, it is Visible *Communication*. Certainly, the defining characteristic of a (full) writing system continues to be its ability to encode speech. However, speech is only one aspect—and perhaps the less important aspect—of oral communication. Just as speech is always also seen through the lens of intonation, eye contact, gestures, and accompanying body language of the speaker,³ writing is framed and often contextualized by its secondary characteristics.⁴

Thus, this article explores an alternative classification that is built on key internal characteristics of a writing system, which is viewed as a set of signs. In this take there are only two main categories of writing systems, namely those that are essentially open and those that are fundamentally closed, though we will see that there is a continuous spectrum in between.

of a given writing system, but the link in terms of the underlying speech has rarely been challenged.

3. “‘Non-verbal behaviour’ refers to actions as distinct from speech. It thus includes facial expressions, hand and arm gestures, postures, positions, and various movements of the body or of the legs and feet,” (Mehrabian, 2007, p. 1). Mehrabian generalizes this to the concept of implicit communication which contains various speech patterns such as intonation, speed etc.

4. Secondary characteristics include the

- relative order of characters in a writing system
- use of glyph variants, e.g.,
 - font variants, e.g., in different fonts
 - different shapes of Chinese Characters in China, Japan and Korea, including the use of historical number signs such as the use of Arabic or Roman numerals
 - font styles such as bold face, italics
 - font sizes
 - usage or not of certain ligatures or ligature groups
- horizontal or vertical arrangement of words and lines (including writing direction)
- use of colours
- general aspects of page layout
- punctuation.

For the concept of secondary characteristics of a writing system cf. Küster (2006).

2. Set of Signs

Most basically, any writing system w_s is a set of signs $\{s_1, \dots, s_n\}$. When applied to a given language only a subset $l_s \subseteq w_s$ will be used.⁵ Some languages such as Japanese combine multiple writing systems.

Throughout the 20th century many l_s contained a small, well-defined set of signs. There was little, if any doubt about the number of signs in l_s . Excepting major historical shifts such as the Russian revolution, which eliminated some letters from the Cyrillic alphabet as used in Russia, or rare successful orthographic reforms, it was impossible to add new signs to l_s or remove existing ones in a manner that would be widely accepted. All l_s used to write European languages fall into this category⁶, but so does e.g., the Hangul script or Hiragana and Katakana⁷. Let's call these w_s closed writing systems.

This is true for most of the long history of these l_s , which have been highly resilient to change once stabilized—famously, even the emperor Claudius with all the might of the Roman empire at his back failed to permanently add three new signs to the Latin alphabet.⁸ Whatever variability existing in manuscript writing was finally eliminated by the introduction of the printing press and its strictly limited number of types.

Other l_s are an obviously finite, but open set of signs. The Chinese script, but also Sumerian cuneiform, Egyptian hieroglyphs and Mayan glyphs are all examples of these. Signs can be added to l_s if a need is felt to do so, other signs fall out of fashion, are only used in specific contexts, or are considered purely historical. While there are compendia attempting to list the signs of l_s , they rarely agree on the complete set of signs, though they typically share a common core set. Let's call these l_s open writing systems.

5. Let us ignore for a moment the question if there is universal agreement that any possible pair of s_i, s_j are different signs or indeed just a variants of a single sign. For the purpose of this argument we can assume that, if in doubt, s_i and s_j will be considered to be two separate signs.

6. Arguably stenography is an exception to this rule, though it is debatable if shorthand is actually a writing system in the first place (in its bid to increase writing speed, shorthand renders speech ambiguously).

7. With Hiragana and Katakana Japanese contains closed writing systems that *could* in principle be used to render Japanese. However, doing so has never been considered a socially acceptable way of writing the language for an educated adult.

8. Suetonius, *Suetoni Tranquillii vita Divi Claudii*, 41. These signs included a letter to represent the consonant reading for U. During the emperor's lifetime these signs were relatively widely used for official inscriptions, but were quickly abandoned after his death.

3. A Continuum of Open and Closed Writing Systems before the Printing Press

Open and closed writing systems have never existed in a vacuum. Depending on the technology used for writing fundamentally closed writing systems would adopt some of the characteristics of open ones and vice versa.

As a case in point, during the middle ages also some of the now closed writing systems could be—and regularly were—extended with new, semi-standardized signs in manuscripts, the *abbreviaturae*.⁹ As the name suggests, they were mainly used to shorten the text and, quite practically, to save valuable writing material such as vellum and to speed up writing. *Abbreviaturae* could normally—though not always reliably—be equated to a string of several canonical signs in w_s .

Similarly, many writing systems had—and in parts still have—the option to use sophisticated ligatures. Ligatures have rarely been considered a mandatory feature for composing texts in a given writing system. As with the *abbreviaturae*, they could and can usually be equated to a string of canonical signs in w_s . Today, Unicode considers ligatures to be solely about choosing the most appropriate glyph to represent a string of characters. However, deciding to use ligatures could and can be an important aesthetic, religious or political statement, e.g., in languages written in the Arabic script¹⁰ where calligraphy is to this day a leading art form. In some cases, the writing of Urdu being an example, these ligatures can become so semantically loaded that they take on characteristics of independent signs for syllables or even words in their own right, resulting in a semi-open writing system, in which the canonical w_s is extended by an in principle open set of ligature signs.

To generalize slightly, semi-open writing systems have maintained a vibrant calligraphic tradition that celebrate also their flexibility and the beauty of glyphs. In fact, a vibrant practice of calligraphy is probably a good indication that the corresponding writing system is open or, at least, semi-open.

Likewise, in Japanese selecting a kanji rather than the corresponding kana(s) can be a necessity to disambiguate the intended meaning between multiple homophones, adding a degree of precision that the spoken language cannot necessarily parallel. However, especially for rarer kanjis outside the standard repertoire, it can also be the author's choice. This choice can characterize a text as more or less sophisticated, scholarly or on the contrary as popular or simple. In Japanese popular literature and films a character's command (or lack of it) of rarer kanjis is

9. Cappelli (1929) remains the standard compendium for *abbreviaturae* in the Latin script.

10. See Küster (2006, 55ff and in particular 57ff) for a fuller discussion.

a regular way to showcase that person as, e.g., a brilliant student or a laggard.

4. Standardization

The counter-tendency is associated with the printing press which with its necessarily limited number of different types almost inevitably brought further standardization to those writing systems most strongly affected by it. McLuhan (1962) rightly considered this launch into the *Gutenberg galaxy* to be a fundamental shift in culture for those living through its impact, and writing—its primary vehicle—was certainly not exempt. Using novel signs continued to be possible in principle, if the typesetter could be convinced to cut suitable types. However, it became a cost factor in a way that it had not been for manuscripts. Under the influence of the printing press quintessentially open writing systems such as Chinese adopted features of in particular the Latin script including numbers, punctuation marks, and direction of writing. The number of signs was often reduced to a core set for everyday use, including formal education, resulting in a semi-closed writing system.¹¹

Typewriters and software in the pre-Unicode age with its very limited character repertoires if anything reinforced this trend for closed writing systems. Handling fundamentally open writing systems—and, often, indeed handling non-Latin writing systems *tout court*—remained beset with many difficulties. Even systems like the original T_EX were mostly invented to overcome their constraints for closed writing systems.

Of course, there have always been exceptions also among closed writing systems. Modernist poetry, including Apollinaire (1918) and Pound (1998), have consciously played with typographic effects to underline the messages of their poems, though tellingly at least Pound was strongly inspired by Chinese poetry in doing so. Concrete poetry thrives on the visual manifestation of a poem. However, none of this has achieved major traction even in poetry, let alone in literature as a whole in any way comparable to Arabic, Chinese, and Japanese calligraphy.

11. For Japanese since 1946 the Tōyō kanji and since 1981 the Jōyō kanji reflect such a semi-closed writing systems. However, they have never intended to contain the whole set of kanji in actual use, but rather the subset that any adult is expected to master.

5. In the Digital Age: Emoji

Unicode / ISO/IEC 10646 and with it OpenType removed many of those constraints by handling all writing systems in a uniform way. Software became—and continues to become—available that allows for a degree of flexibility almost impossible to achieve with the printing press. In addition, Unicode allows to easily mix characters from almost arbitrary writing systems in a single text, and be reasonably sure that most systems would be able to render it without too many issues.

However, by creating a single, open, monumental meta-writing system, Unicode might involuntarily have reversed the arrow of influence—open writing systems start to influence closed ones, largely via the unlikely vehicle of emoji. “Emoji are ‘picture characters’ originally associated with cellular telephone usage in Japan [...] Emoji are often pictographs—images of things such as faces, weather, vehicles and buildings, food and drink, animals and plants—or icons that represent emotions, feelings, or activities.”¹²

Some emoji such as 😊 or 😍 symbolize emotions (emoticons). Others such as 🐼, 🍷 or 🍀 are depictions of the respective objects, though some as 🍀 can also be used as metaphors for abstract concepts (here luck). Other emoji such as 📅 always represent abstractions (day or date in this case).

Emoji can be used in many contexts: as pure mood markers in a role similar to punctuation, as a means to stress the text, as replacement for individual words, concepts, or as a semantic marker.

Let me illustrate these phenomena by some recent, more or less randomly selected tweets, all from official sources:

—
67% of #Luxembourg residents believe integration of most immigrants is successful. Immigration from outside #EU is seen as an opportunity by twice as many people as opposed to a problem. Infos on integration of immigrants from 🇱🇺 perspective ➡
blob:ec.europa.eu/6dd004ba-9574- ...

The Luxembourgian flag obviously stands for the corresponding toponym, the arrow for “see also”.¹³

The Erasmus+ programme uses emoji purely for emphasis.¹⁴

12. Unicode Consortium (2016), cf. also Küster (2016).

13. https://twitter.com/Yuriko_Backes/status/985018574254346240, retrieved on 2018-04-14. The colours of the hashtags and links are, however, a phenomenon of the chosen Twitter client.

14. <https://twitter.com/EUErasmusPlus/status/984868830873899008>, retrieved on 2018-04-14.

What do #social rights mean to you?

📷 Show us in a photo for the chance to attend this year's European Youth Event in Strasbourg and win some amazing prizes, including two Interrail passes! 🙌

Enter your photo now 📲 woobox.com

[/ugd2pa](#)

[#MySocialRights](#)

A particularly nice example illustrating multiple of these usage modes comes from a Louisiana police department, the text of which could be read “It is time for those 9pm checks. Start by removing those valuables such as your phone, watch, and computer. Don’t forget about locking the car door. Last but not least, close the doors of your home and set the alarm [emphasis]”¹⁵

It’s almost time for those 9pm checks. Start by removing those valuables 📱 ⌚ 💻. Don’t forget about 🔑 the 🚗 door. Last but not least, close the 🏠 doors 🚪 and set the alarm 🚨. #9PMRoutine

Even if the linguistic rendering in cases like 🏠 might be underdetermined—it could be read as “the house,” “your house,” “your home” etc.—, the context usually disambiguates the intended message.¹⁶ The objective of the writer is not necessarily to encode a precise utterance, but rather to transmit a message - and in spite of some ambiguous readings the semantics are clear.

Emoji are a very recent phenomenon, having become a global phenomenon only over the last few years. After having been popularized in the late 1990s in Japan,¹⁷ they owe their worldwide success to their incorporation into the Unicode standard and subsequently into all major operating systems. Today, emoji are regularly used in email exchanges—often, but not necessarily as emoticons—, but are omnipresent in short messaging, tweets and similar short, informal means of communication to the point that iOS by default now offers a second keyboard just for emoji and Android devices systematically add emoji among their suggested word completions when typing.¹⁸ Even high profile newspapers

15. <https://twitter.com/CreveCoeurPD/status/959254450769670144>, retrieved 2018-02-02

16. Cf. also Dürscheid and Siever (2017) for corresponding examples in German.

17. A very readable, though not academic history of emoji was published in <https://www.theverge.com/2013/3/4/3966140/how-emoji-conquered-the-world>

18. Cf. on this also Dürscheid and Siever (ibid.), passim.

such as the *New York Times* play with the omnipresence of emoji,¹⁹ though as of yet only in their more experimental and youth-oriented features.

Emoji can shorten a text, convey a speaker's intentions, create subtexts, and be funny. Throughout the world they frequently appear on clothing, in popular art, even on food in a way that signs of closed writing systems rarely do. The use of emoji seems to be growing continuously and gain currency also in more conventional forms of publications.

6. Mixing Writing Systems

6.1. Mixing Scripts

Up until very recently mixing scripts in Western languages was more or less the domain of scholarly or scientific writing. Scholars need to quote citations written in other scripts such as Greek or Hebrew. Scientists use Greek or (rarely) Hebrew characters, mainly in mathematical or physical formulas. Both scenarios were traditionally the domain of academic publishers and until quite recently only supported by specialized software.

Emoji are different. Like in Japan, in the West emoji are primarily found in informal writing, especially messaging, but they start to encroach on more traditional publication channels.

This hurdle has always been lower for languages such as Japanese which is traditionally written using a number of scripts. The anime advertisement in Figure 1 targeted at teenagers showcases on one page kanji, hiragana, katakana, Latin, various punctuation signs, and even a Greek character, μ (the name of the band).

6.2. Unicode: A Single, Open, Monumental Meta-Writing System

Emoji and in general more regular mixing of writing systems would not have been possible without a truly universal character set that can encode a number of characters that is likely to suffice for all characters that have been and will ever be invented.²⁰

In addition, it is almost universally supported across operating systems, types of devices and software.

19. An example of several: Schulten (n.d.). This article appeared in the New York Times learning network targeted at educational institutions.

20. Technically, of course, Unicode / ISO/IEC 10646 is a 32-bit writing system, making it a very large, but still finite writing system.



FIGURE 1. <http://www.lovelive-anime.jp/otonokizaka/>, retrieved on 2019-05-12

7. Summary and Outlook

New emoji are created on a regular basis, though they don't necessarily all see widespread adoption. Other emoji go out of fashion, many more are used without a universally agreed interpretation. This does not seem to impede their growing popularity. While some writers have experimented with emoji-only texts, these have not received much traction. However, mixed texts where characters from existing *l*_s intermingle with emoji have become the norm in some communication channels, notably in messaging and on social media.

Contrary to *abbreviaturae* and ligatures, emoji have the potential to fundamentally open up closed writing systems. Only time will tell if they will remain a short-lived phenomenon or if they will change the character of closed writing systems throughout the world.

References

- Apollinaire, Guillaume (1918). *Calligrammes. Poèmes de la Paix et de la Guerre (1913-1916)*. Paris: Mercure de France.
- Cappelli, Adriano (1929). *Lexicon Abbreviaturarum. Dizionario di Abbreviature Latine ed Italiane*. 3rd ed. Milano: Hoepli.

- Daniels, Peter T. (1996). "Grammatology: Introduction". In: Daniels, Peter T. and William Bright. *The World's Writing Systems*. Oxford: Oxford University Press, pp. 1–2.
- DeFrancis, John (1989). *Visible Speech. The Diverse Oneness of Writing Systems*. Honolulu: University of Hawai'i Press.
- Dürscheid, Christa and Christina Margrit Siever (2017). "Jenseits des Alphabets - Kommunikation mit Emojis". In: *Zeitschrift für germanistische Linguistik* 45.2, pp. 256–285.
- Gelb, Ignace J. (1963). *A Study of Writing*. 2nd ed. Chicago: University of Chicago Press.
- Küster, Marc Wilhelm (2006). *Geordnetes Weltbild*. Tübingen: Niemeyer.
- (2016). "Writing beyond the letter". In: *Tijdschrift voor Media-geschiedenis [Journal for Media History]* 19.2, pp. 1–17.
- McLuhan, Marshall (1962). *The Gutenberg Galaxy. The Making of Typographic Man*. Toronto: University of Toronto Press.
- Mehrabian, Albert (2007). *Nonverbal Communication*. London, New York: Routledge.
- Pound, Ezra (1998). *The Cantos of Ezra Pound*. London: Faber & Faber.
- Schulten, Katherine (n.d.). "Emojis". <https://www.nytimes.com/2019/05/10/learning/emojis.html>.
- Taylor, Isaac (1883). *The Alphabet. An Account of the Origin and Development of Letters. Semitic Alphabets*. London: Kegan Paul, Trench, & Co.
- Unicode Consortium (2016). "FAQ—Emoji and Dingbats". http://unicode.org/faq/emoji_dingbats.html.

The History of the Graphematic Foot in English and German

Martin Evertz

Abstract. Suprasegmental graphematics holds that there are units in alphabetical writing systems comprising more than one segment. While units such as the graphematic syllable and the graphematic word seem to be well established, the graphematic foot was only recently proposed. This paper provides further insights into this unit by discussing diachronic data from English and German.

There are two phenomena that make the graphematic foot especially visible: graphematic geminates in English and German and the silent <e> in English. Both phenomena coded segmental information in earlier stages of the languages, i.e., spelling geminates coded phonological geminates and the final -e in English coded schwa. At some time, phonological geminates in both languages and the word-final schwa in English disappeared. That rendered the original functions of these spelling devices obsolete. However, instead of vanishing, graphematic geminates and the final -e acquired new functions connected to the graphematic foot.

The phonological segments, which were coded by the discussed phenomena, developed because of suprasegmental conditions: geminates and the word-final schwa played a major role in the development of the vowel quantity systems of both languages, which is connected to syllable and foot structure. In today's systems, the graphematic foot bidirectionally corresponds to the phonological foot and thus helps the reader to gain information about the phonological foot and syllable structure of a word.

This new diachronic approach may not only enhance our understanding of the unit graphematic foot but it may also help to understand how and why suprasegmental units developed in writing systems in the first place.

1. Introduction

In traditional writing system research, written language is analysed as a linear sequence. Contrary to this view, suprasegmental graphematics holds that there are units in alphabetical writing systems comprising more than one segment, which are organized in a hierarchy parallel to the phonological hierarchy (cf. Evertz and Primus 2013; Evertz

Martin Evertz
University of Cologne
martin.evertz@uni-koeln.de

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1.
Fluxus Editions, Brest, 2019, p. 27–40. <https://doi.org/10.36824/2018-graf-ever>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

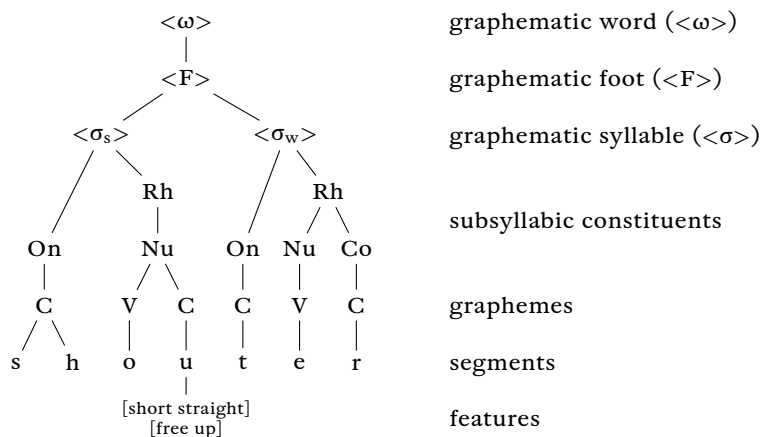


FIGURE 1. The graphematic hierarchy (cf. Evertz and Primus 2013; Evertz 2018)

2018). Moreover, suprasegmental graphematics claims that written language bidirectionally corresponds to spoken language (cf. Primus 2003; U. Domahs and Primus 2015). The units which make up the *graphematic hierarchy* are available in spoken and written language.

The existence and relevance of some of these units are quite uncontroversial; for example, there is no doubt that the graphematic word is a relevant unit (at least in writing systems of languages such as English and German). Other units, e.g., the graphematic syllable, are widely acknowledged in the literature (cf. e.g., Butt and Eisenberg 1990; Roubah and Taft 2001; Rollings 2004; Primus 2003; F. Domahs, Bleser, and Eisenberg 2001). The unit which will be the focus of this paper, the graphematic foot, was only recently proposed for German (Primus 2010) and English (Evertz and Primus 2013; Evertz 2016; Ryan 2017; Evertz 2018). The whole graphematic hierarchy is shown in Fig. 1.

The research on the graphematic foot so far has taken a synchronic perspective. This paper is the first attempt at shedding some light on the history of the graphematic foot. This might provide some insights on how suprasegmental units come into being and may be the foundation for explaining some until now only poorly understood phenomena.

The remainder of this paper is organized as follows: In order to lay a foundation for the discussion of the graphematic foot, I will discuss some phonological preliminaries. After that I will discuss how the graphematic foot developed. I will argue that the development of the obligatory branching nucleus in stressed syllables was one of the key changes in the prosodic systems of English and German that impacted the development of the graphematic foot. The paper closes with a short

conclusion, in which I will briefly summarise the findings presented in this paper.

2. Phonological Preliminaries

In order to understand the phonological changes in the history of English and German some phonological facts have to be established first.

The theory of prosodic phonology (e.g., Selkirk 1980; 1981; Nespor and Vogel 1986) holds that speech is arranged into hierarchically organised constituents. These constituents form the domains for phonological rules or constraints, which are joined together into a hierarchical structure known as the prosodic or *phonological hierarchy*. Most theories agree that the phonological hierarchy contains at least the syllable, the foot, the phonological word and one or more constituents above the word (cf. Shattuck-Hufnagel and Turk 1996, for a comparison of the constituent inventories of some of the most influential theories). In this paper, we will focus on the syllable and the foot.

Under minimal assumptions, the principal subparts of the syllable are the syllable peak and the two margins, which can be called onset and coda. The syllable peak contains the most sonorous segment, where sonority is an abstract property of a segment (Zec, 2007). It is defined as the (sole) sonority peak of a syllable and represented as a structural position V. V does not necessarily dominate a vowel. In languages such as English and German, the V-slot can also be occupied by liquids and nasals in unstressed syllables. Non-peak positions are denoted by C and must not necessarily dominate a consonant; this is, for instance, the case in the representation of diphthongs, in which the second vowel of the diphthong is dominated by C (cf. Clements and Keyser 1983).

A non-linear syllable model such as the CV-model can represent vowel opposition between long/ tense and short/lax vowels in languages such as contemporary English and German by the association of long/tense vowels with two structural positions while short/lax vowels are associated with one structural position, cf. Fig. 2a in which the vowel of the first syllable is dominated only by V, while in Fig. 2b the vowel of the first syllable is dominated by V and C. Note that the structural representations of *filler* and *poker* in Fig. 2 hold for German and English.¹

In modern English, some tense vowels are realised as diphthongs in many varieties, including Received Pronunciation and General American English (cf. Giegerich 1992, pp. 44–47). A diphthong as in the received pronunciation of *poker* is analysed and represented as an under-

1. In Standard German, the last syllable of *Poker* and *Filler* is open and ends in [ɐ]; in American English, both words end in [ə]. The illustrations in Fig. 2 are approximations.

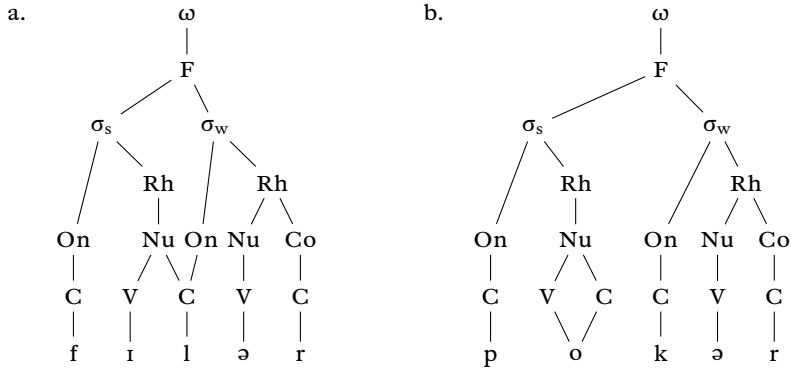


FIGURE 2. Phonological representation of *filler* and *poker* in modern English and German, cf. Evertz and Primus (2013, p. 4)

lying tense vowel, as shown in Fig. 2b. Tense vowels and diphthongs alternate, as in *line* – *linear*, *provoke* – *provocative* and *bathe* – *bath*. The phonetic correlate of the vowel contrast under discussion is a matter of debate and the terminology varies considerably (e.g., tense – lax, long – short, free – checked). Due to the structural property of tense vowels and diphthongs to occupy two structural positions, I will call them *binary* vowels. Lax vowels occupy one structural position and, hence, are *unary*.

In addition to the CV-tier, most phonologists assume that there is a richer structure with mediate constituents between the CV-tier and the σ -node. I will adopt a syllable structure model in which a syllable necessarily comprises a rhyme (Rh) which dominates a nucleus (Nu) that in turn dominates the V-position. Optional subsyllabic constituents are the onset (On) and the coda (Co), cf. Fig. 2.

An important observation for contemporary English and German is that in both languages stressed syllables may never end in a unary vowel. A stressed syllable or even a monosyllabic word like */pɪ/ or */pɛ/ is ill-formed in modern English and German. This property of stressed syllables in English and German can be accounted for by a syllable structure constraint demanding that the nucleus of a stressed syllable is obligatory branching (cf. Becker 1996). According to Wiese (2000, pp. 46–47) all full, stressed or unstressed, syllables have a branching nucleus that dominates V and C. A similar restriction is formulated by Giegerich (1992, p. 182) in terms of a branching rhyme. We will see that this constraint began to develop in Old English and Old High German. I will argue that the development of the branching nucleus is the key in understanding how the graphematic foot developed in English and German.

The next higher unit, the phonological foot, is defined as a sequence of one or more syllables, in which exactly one syllable is the head of the foot, i.e., stressed/strong. In German and English, the default foot pattern is trochaic. In other words, feet in English and German are by default head-initial. For a recent overview and comparison of the phonological foot in English, German and Dutch see U. Domahs, Plag, and Carroll (2014).

As previous work on the graphematic foot shows, phonological structures and constraints discussed here have close correspondents in graphematic structures and graphematic constraints (cf. Evertz and Primus 2013; Fuhrhop and Peters 2013; Evertz 2018). It is important to understand, however, that graphematic structures and constraints are not derived from phonology. In this model, phonology and graphematics are two interdependent systems connected by bidirectional correspondences, all graphematic constraints are motivated independently on graphematic grounds (cf. Evertz 2018; Evertz and Primus 2013).

3. Before and during the Rise of the Branching Nucleus

In this paper, we will examine time periods of English and German before and after the development, or rise, of the branching nucleus in the prosodic systems of these languages. The time periods under discussion are Old English (OE; ca. 450 to 1150 CE) and Old High German (OHG; ca. 700 to 1050 CE), Middle English (ME; ca. 1150 to 1500), modern English (from ca. 1550 on) and modern German (from ca. 1650 on). The rise of the branching nucleus began in the middle periods of the languages discussed here.

3.1. Phonological Realisation of Gemination and Final -e

In Old English (3.1) and Old High German (3.1) geminate (long) consonants contrast with single (short) consonants. The following minimal pairs thus demonstrate that gemination in OE and OHG was relevant on a phonemic level and that it was phonological distinct from single consonants, cf. Britton (2012) and Simmler (2000).

(1) *wike* /k/ 'week' vs. *wikke* /k:/ 'wicked'; *sune* /n/ 'son' vs. *sunne* /n:/ 'sun'

(2) *miti* /t/ 'thereby' vs. *mitti* /t:/ 'middle'; *filu* /l/ 'much' vs. *fillu* /l:/ 'I beat'

Final -e (schwa) developed in Middle English due to vowel reduction and was not mute but contrasted with other vowels, cf. (3.1), Minkova (1991).

- (3) *bode* ‘message’ vs. *bodi* ‘body’; *dule* ‘devil’ vs. *duly* ‘truly’

The examples provided here show two things: first, in earlier stages of English and German, there is a contrast of long and short consonants, and this contrast is marked by graphematic gemination, i.e., by doubled letters. Second, the final -e in English used to correspond to a vowel.

3.2. Gemination in OE and OHG

In the late stages of OE and OHG, the quantity and stress system of both languages began to change. One of the major developments was *vowel shortening*. Long vowels and diphthongs in strong syllables were shortened especially before geminates, before three consonants, and before groups of two consonants in polysyllabic forms if at least two unstressed syllables followed (Lahiri, Riad, and Jacobs, 1999, p. 347). Thus, it seems that vowel shortening often happened in order to avoid overlong syllables.

Vowel shortening could also occur in words which do not fit in the description above, for instance in words like *blāder* ‘ladder’. If in a word like this the vowel is shortened, this shortening could be compensated by the gemination of the consonant that immediately follows that vowel. Thus, vowel shortening could trigger gemination (Hickey, 1986), see (3.2).

- (4) Development in late OE
- a. *blāder* → *bladder* ‘ladder’
 - b. *fōder* → *fodder* ‘fodder’

Let us have a look at the syllable structure of the words in (3.2). A word like *blāder* consists of two syllables. The vowel in the first syllable occupies two structural positions. In other words, the syllable nucleus is branching. Due to vowel shortening, the vowel in the first syllable becomes short and occupies only one structural position; the second structural position that used to be occupied by the long vowel becomes free. This shortening is compensated by the geminate: the geminate occupies the structural position that became free.

This leads to the conclusion that the second structural position of the nucleus in a stressed syllable must not be free, it must be occupied by a vowel (either a long vowel or the second element of a diphthong) or by a consonant. In other words, the nuclei of stressed syllables became obligatory branching.

The phonological structure of words with a geminated consonant can be reconstructed like in Fig. 3a. (adapted from the phonological structure of gemination in contemporary languages, cf. Davis 2011).

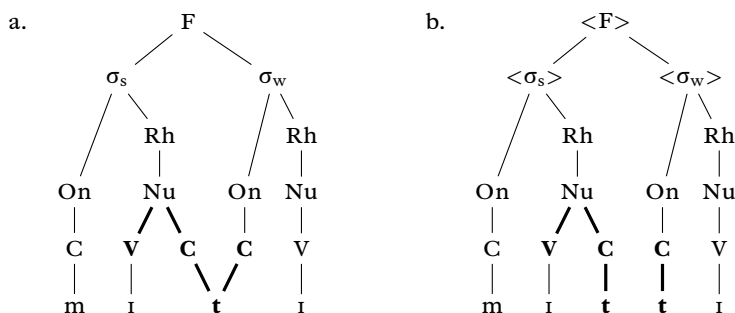


FIGURE 3. Phonological foot structure of words with geminates in OE and OHG (a.) and their graphematic structure (b.), example OHG *mitti* 'middle'

The positions of the phonological geminates within the syllable structure is identical in OHG and OE: the geminate occupies the last position of the rhyme of the first syllable and simultaneously the first position of the onset of the following syllable. The geminate is associated with two skeletal positions and is thus long. In the graphematic representation, it seems that a single letter cannot be associated with two structural positions. This is in line with findings pertaining the representation of ambisyllabicity in modern German (cf. Eisenberg 1989, p. 82; Primus 2003, p. 35). The geminated consonant is thus indicated by a geminated (doubled) letter. These letters are also associated with two skeletal positions, see Fig. 3b.

3.3. Final -e in ME

We have seen in the previous section that from OE and OHG on, English and German developed obligatorily branching nuclei in stressed syllables. In other words, at least the structural position dominated by the syllable peak of a stressed syllable and the immediately following position must not be empty but associated with a segment.

The lengthening process which took place in the middle periods of English and German commonly dubbed *open syllable lengthening* fits into the development of branching nuclei in stressed syllables. In open syllable lengthening, short vowels occurring in open syllables were lengthened (Lahiri, Riad, and Jacobs, 1999, p. 350). At the same time, a process commonly dubbed *vowel reduction* reduced unstressed full vowels at the end of words to schwa (Minkova, 1991), see (3.3).

- (5)
- | | OE | | ME | |
|----|-------------|---|-------------|--------|
| a. | <i>wūdu</i> | → | <i>wōde</i> | ‘wood’ |
| b. | <i>nāme</i> | → | <i>nāme</i> | ‘name’ |
| c. | <i>nōsu</i> | → | <i>nōse</i> | ‘nose’ |

From a structural perspective this means that the empty position after the syllable peak is filled by associating this position with the vowel, i.e., by lengthening it. Fig. 4 is a reconstruction of the phonological structure of ME *name*. Note that the final schwa opens the first syllable by taking [m] as onset. The graphematic representation of *name* is identical to its phonological counterpart.

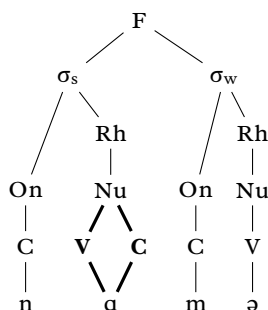


FIGURE 4. Phonological foot structure of words with final -e and one intervocalic consonant, example ME *name*

3.4. Cause and Effects in the Phonological Systems

As we have seen in the previous sections, gemination and open syllable lengthening were caused by a reorganisation of the prosodic systems of English and German, especially in terms of quantity and stress. From a structural perspective, one of the major changes was the rise of the branching nucleus, i.e., the nucleus of stressed syllables became obligatory branching.

Gemination and final schwa were coded in a transparent way: phonological gemination was coded by graphematic gemination, i.e., by doubling the letter that corresponds to the geminated consonant. Since the final -e corresponds to a vowel, schwa, it was coded by <e>.

In the middle periods, phonological geminates disappeared in English and German and the final -e (schwa) in English – but not in German

– became mute.² After the disappearance of geminates and the muting of the final -e, the doubled consonant and the final -<e> became obsolete. But instead of vanishing, these spelling devices acquired new functions connected with the graphematic foot, as I will show in the following sections.

4. After the Rise of the Branching Nucleus

4.1. Ambisyllabicity (De-)coding

Because stressed syllables developed branching nuclei, a single consonant adjacent to two single vowels (the first one being short and in the stressed syllable) is ambisyllabic, cf. Fig. 5a (Giegerich 1992, pp. 170–172; Wiese 2000, pp. 46–47; McMahon 2001, pp. 111–112). An ambisyllabic consonant is a consonantal segment that simultaneously belongs to the rhyme of one syllable and to the onset of the immediately following syllable. Early influential accounts promoting this concept include Kahn (1976) and Gussenhoven (1986) for English, and Vennemann (1982) for German.

On first glance, gemination and ambisyllabicity might appear quite similar. Both phenomena involve consonants with ambiguous associations to syllables. But while geminated consonants occupy two structural positions where the first position belongs to the nucleus of one syllable and the second position belongs to the onset of a following syllable, an ambisyllabic consonant is associated with one structural position only. This position is simultaneously dominated by the nucleus of one syllable and the onset of a following syllable. On the surface, this difference can be perceived as a difference in quantity: geminated consonants are long while ambisyllabic consonants are not.

Due to geminate loss, the earlier geminate (de-)coding (cf. Fig 3) became obsolete. But instead of vanishing, the geminate (de-)coding was reinterpreted as ambisyllabicity (de-)coding by the graphematic system, cf. Fig. 5b and Fig. 3b.

Note that in modern English and modern German, this system is obscured in some cases. As Evertz and Primus (2013, p. 9) point out, there are independent constraints which can block the gemination of some consonant letters. For instance, complex graphemes (such as <sh> in English or *sch* in German) or other letters such as <v> cannot be geminated. Words such as *navvy* and *skivvy* are marginal (cf. Cook 2004, p. 60), but they show the tendency to violate a highly ranked constraint ('do not geminate <v>') in order to conform to the model presented here

2. These changes can be as well attributed to the establishment of the current syllable and foot structure, cf. Britton (2012).

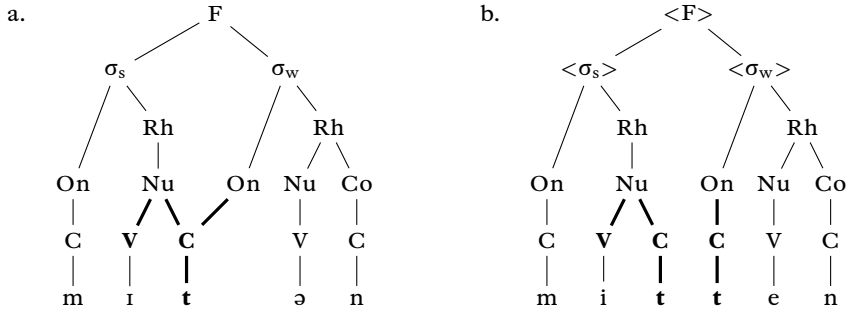


FIGURE 5. a.: Ambisyllabicity in Engl. and Ger.; b.: graphematic gemination, example *mitten*, Engl. a type of glove, Ger. ‘(in the) middle’

(cf. also Ryan 2010, p. 31). Words such as *give* and *dive* are opaque with respect to the vowel contrast under discussion.

4.2. Final -e in Modern English

In late middle English and early modern English, final -e lost its phonological correspondent (schwa). The graphematic structure for final schwa leading to vowel lengthening (see Fig. 4) persisted and was reinterpreted as a sign of vowel length, i.e., a vowel in a branching nucleus.

Structurally speaking, the final <e> constitutes a graphematic syllable, which in turn constitutes a graphematic foot together with the preceding syllable. Because the nucleus in a strong syllable branches, a single vowel consonant in an open graphematic syllable is interpreted to be associated to two structural positions. A reader thus can infer that this vowel letter corresponds to a binary vowel.

Although the final -e is mute, it *visually* opens the first syllable of words like <name>, Fig. 6b. Because of that, the reader can infer that the corresponding phonological syllable is branching, Fig. 6a.

It has to be noted, however, that this model does not hold for every occurrence of final -e in today’s English. Evertz and Primus (2013, p. 9) point to following exceptional patterns:

- i. <o+Nasal+e> for a unary vowel: *done, one, come, some*
- ii. <e> after <s> distinguishing stem final from inflectional <s>: *goose, mouse, cheese, dense, tense*. This kind of <e> does not disambiguate the phonological value of the first vowel.
- iii. idiosyncratic cases: *camel, belle, tulle*

Some instances in which this model does not hold are explicable by their non-native origin: for instance, the word *belle* with a unary vowel and a

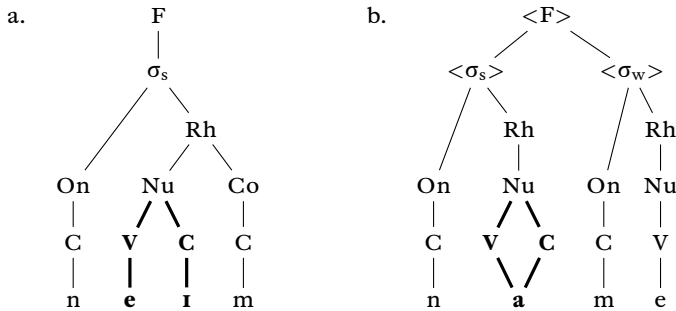


FIGURE 6. Phonological (a.) and graphematic (b.) foot structure of the word *name*

superfluous mute <e> and *tulle* with a binary vowel and an irregular c-gemination are explicable by their Modern French origin (cf. Venezky 1999, p. 86).

5. Conclusion

This paper presents some insights of how the graphematic foot developed in English and German. The graphematic foot is considered to be a suprasegmental unit in the writing systems of English and German that bidirectionally corresponds to the phonological foot.

There are two phenomena in the writing systems of today's English and German that make the graphematic foot especially visible, graphematic geminates (i.e., doubled consonant letters) and silent <e> in English. Originally, graphematic geminates and the final <e> were coding segmental information: graphematic geminates coded phonological geminates (i.e., long consonants) and word-final <e> coded word-final schwa. Phonological geminates and word-final schwa in turn developed because of suprasegmental conditions: they played a major role in the reorganisation of the prosodic systems of both languages (especially in terms of quantity and stress).

During the reorganization of the prosodic systems of English and German, phonological geminates disappeared and final -e became mute in English. This rendered the connected spelling devices, i.e., graphematic geminates and word-final <e>, obsolete. But instead of vanishing, graphematic geminates and final <e> acquired new functions.

In middle English and middle German the nuclei of stressed syllables became obligatory branching, this means that the syllable peak and structural position immediately following must not be empty. It follows that an open stressed syllable can never have a short vowel. This leads

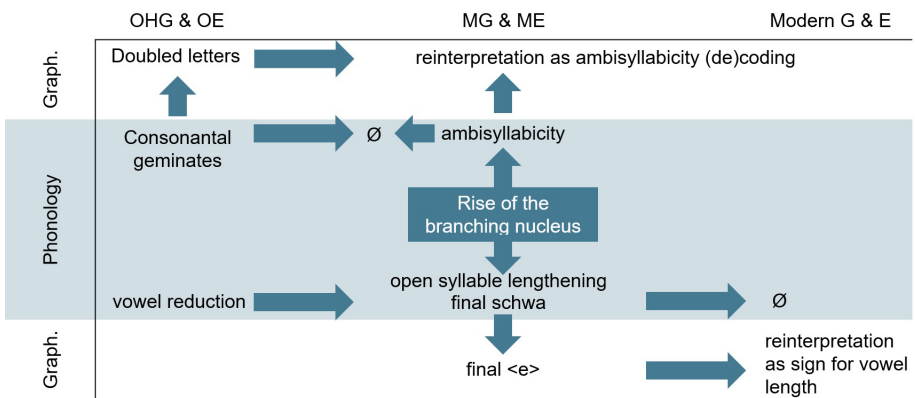


FIGURE 7. Summarizing model of the development of two of the most prominent phenomena connected to the graphematic foot

to ambisyllabicity in words in which a single consonant is adjacent to a short vowel in a stressed syllable and another syllable peak. Graphematic geminates that used to correspond to phonological geminates were reinterpreted to (de)code ambisyllabic consonants.

Silent <e> on the other hand is used to (de)code vowel quantity. Although the final <e> is mute, it visually opens graphematic syllables. Because the nucleus of a strong syllable (in phonology and in graphematics) is branching, a single vowel letter in an open graphematic syllable that is the head of a graphematic foot is interpreted as (de)coding a long vowel.

In short, after phonological geminates disappeared and final -e became mute, their graphematic correspondents, graphematic geminates and the final <e>, acquired new functions connected to the graphematic foot, cf. Fig. 7 for a summary.

References

- Becker, Thomas (1996). "Zur Repräsentation der Vokallänge". In: *Zeitschrift für Sprachwissenschaft* 15, pp. 3–21.
- Britton, Derek (2012). "Degemination in English, with Special Reference to the Middle English Period". In: *Analysing Older English*. Ed. by David Denison et al. Cambridge: Cambridge University Press, pp. 233–243.
- Butt, Matthias and Peter Eisenberg (1990). "Schreibsilbe und Sprechsilbe". In: *Zu einer Theorie der Orthographie*. Ed. by Christian Stetter. Tübingen: Niemeyer, pp. 33–64.

- Clements, G. N. and S.J. Keyser (1983). *CV Phonology*. Cambridge, MA: MIT Press.
- Cook, Vivian (2004). *The English Writing System*. London: Routledge.
- Davis, Stuart (2011). "Geminates". In: *The Blackwell Companion to Phonology*. Ed. by Marc van Oostendorp et al. Malden, MA, Oxford: Wiley-Blackwell, pp. 837–859.
- Domahs, Frank, Ria de Bleser, and Peter Eisenberg (2001). "Silbische Aspekte segmentalen Schreibens – neurolinguistische Evidenz". In: *Linguistische Berichte* 185, pp. 13–30.
- Domahs, Ulrike, Ingo Plag, and Rebecca Carroll (2014). "Word Stress Assignment in German, English and Dutch: Quantity-Sensitivity and Extrametricality Revisited". In: *Journal of Comparative Germanic Linguistics* 17.1, pp. 59–96.
- Domahs, Ulrike and Beatrice Primus (2015). "Laut – Gebärde – Buchstabe". In: *Sprache und Wissen*. Ed. by Ekkehard Felder and Andreas Gardt. Berlin: de Gruyter, pp. 125–142.
- Eisenberg, Peter (1989). "Die Schreibsilbe im Deutschen". In: *Schriftsystem und Orthographie*. Ed. by Peter Eisenberg and Hartmut Günther. Tübingen: Niemeyer, pp. 57–84.
- Evertz, Martin (2016). "Minimal Graphematic Words in English and German. Lexical Evidence for a Theory of Graphematic Feet". In: *Written Language & Literacy* 19.2, pp. 189–211.
- (2018). *Visual Prosody—the Graphematic Foot in English and German*. Berlin: De Gruyter.
- Evertz, Martin and Beatrice Primus (2013). "The Graphematic Foot in English and German". In: *Writing Systems Research* 5.1, pp. 1–23.
- Fuhrhop, Nanna and Joerg Peters (2013). *Einführung in die Phonologie und Graphematik*. Stuttgart: J. B. Metzler.
- Giegerich, Henz J. (1992). *English Phonology: An Introduction*. Cambridge: Cambridge University Press.
- Gussenhoven, Carlos (1986). "English Plosive Allophones and Ambisyllability". In: *Gramma* 10, pp. 119–141.
- Hickey, Raymond (1986). "Remarks on Syllable Quantity in Late Old English and Early Middle English". In: *Neuphilologische Mitteilungen* 87, pp. 1–7.
- Kahn, Daniel (1976). "Syllable-Based Generalizations in English Phonology". PhD thesis. MIT.
- Lahiri, Aditi, Tomas Riad, and Haike Jacobs (1999). "Diachronic Prosody". In: *Word Prosodic Systems in the Languages of Europe*. Ed. by Harry van der Hulst. Berlin: de Gruyter, pp. 335–422.
- McMahon, April (2001). *An Introduction to English Phonology*. Edinburgh: University Press.
- Minkova, Donka (1991). *The History of Final Vowels in English. The Sound of Muting*. Berlin: de Gruyter.

- Nespor, Marina and Irene Vogel (1986). *Prosodic Phonology*. Fordrecht: Foris.
- Primus, Beatrice (2003). "Zum Silbenbegriff in der Schrift-, Laut- und Gebärdensprache – Versuch einer mediumübergreifenden Fundierung". In: *Zeitschrift für Sprachwissenschaft* 22, pp. 3–55.
- (2010). "Strukturelle Grundlagen des deutschen Schriftsystems". In: *Schriftsystem und Schriffterwerb: linguistisch – didaktisch – empirisch*. Ed. by Ursula Bredel, Astrid Müller, and Gabriele Hinney. Tübingen: Niemeyer, pp. 9–45.
- Rollings, Andrew G. (2004). *The Spelling Patterns of English*. Munich: Lincom Europa.
- Roubah, Aïcha and Marcus Taft (2001). "The Functional Role of Syllabic Structure in French Visual Word Recognition". In: *Memory & Cognition* 29, pp. 373–381.
- Ryan, Des (2010). "Kre-8-iv Spell!nk: Why Constructed Homophony Is Key to Understanding Patterns of Orthographic Change". MA thesis. Edinburgh University.
- (2017). "Principles of English Spelling Formation". PhD thesis. Trinity College Dublin.
- Selkirk, Elisabeth O. (1980). "The Role of Prosodic Categories in English Word Stress". In: *Linguistic Inquiry* 11, pp. 563–605.
- (1981). "On the Nature of Phonological Representation". In: *The Cognitive Representation of Speech*. Ed. by John Anderson, John Laver, and Terry Myers. Amsterdam: North Holland, pp. 379–388.
- Shattuck-Hufnagel, Stefanie and Alice E. Turk (1996). "A Prosody Tutorial for Investigators of Auditory Sentence Processing". In: *Journal of Psycholinguistic Research* 25.2, pp. 193–247.
- Simmler, Franz (2000). "Phonetik und Phonologie, Graphetik und Graphematik des Althochdeutschen". In: *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. HSK 2.2*. Ed. by Werner Besch et al. Berlin: de Gruyter, pp. 1155–1170.
- Venezky, Richard L. (1999). *The American Way of Spelling: The Structure and Origins of American English Orthography*. New York, London: The Guilford Press.
- Vennemann, Theo (1982). "Zur Silbenstruktur der deutschen Standardsprache". In: *Silben, Segmente, Akzente*. Ed. by Theo Vennemann. Tübingen: Niemeyer, pp. 261–305.
- Wiese, Richard (2000). *The Phonology of German*. 2nd ed. Oxford: Oxford University Press.
- Zec, Draga (2007). "The syllable". In: *The Cambridge Handbook of Phonology*. Ed. by Paul de Lacy. Cambridge: Cambridge University Press, pp. 161–194.

Graphemic Methods for Gender-Neutral Writing


Yannis Haralambous & Joseph Dichy

Abstract. In this paper we present a model and a classification of graphemic gender-neutral writing methods, we explore current practices in French, German, Greek, Italian, Portuguese and Spanish languages, and we investigate interactions between gender-neutral writing forms and regular expressions.

1. Introduction: The General Issue of, and behind, Gender-Neutral Writing

The issue behind gender-neutral writing is that of the representation of inter-gender relations carried by languages. What is at stake is the representation of equality, or not, between genders. The issue is also referred to as “inclusive writing,” which apparently refers to human rights, but does not cover all cases, i.e., lesbian, gay, bisexual and trans-gender persons. The term “Gender-Neutral Writing” is, in fact, both clearer and more inclusive.

Generally speaking, languages are conservative, if not archaic. A significant example is that of the idea of time, which is traditionally represented as a dot sliding along a straight line in a continuous movement. This has been a philosophical image of time since ancient Greek philosophers and throughout the Middle Ages both in Arabic and European philosophy, but can no longer be considered as a valid representation after 20th century existentialist philosophers and Heidegger’s *Sein und Zeit*.

Yannis Haralambous  0000-0003-1443-6115
IMT Atlantique & LabSTICC UMR CNRS 6285
Brest, France
yannis.haralambous@imt-atlantique.fr

Joseph Dichy
Professor of Arabic linguistics
Lyon, France
joseph.dichy@yahoo.fr

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1.
Fluxus Editions, Brest, 2019, p. 41–89. <https://doi.org/10.36824/2018-graf-hara2>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

FIGURE 1. Use of gender-neutral writing on the Web page of IMT Atlantique as a means to attract engineering students of both genders (“Bring out the engineer_{gn} in you!”)

This is even more true considering Einstein’s relativity or any representation in modern physics. Nevertheless, the dot-moving-along-a-line representation remains efficient in the study of time and aspect in natural languages, because language structures reflect the naive or archaic view of human language users.

When it comes to gender, needless to say, the representation of inter-gender relations in linguistic lexical and grammatical structures as well as in discourse also remains archaic. Inclusive writing proposes to change the representation of genders in a way that puts forward equality between women and men. It nevertheless raises many questions, some of which are of a social and/or ideological nature, including educational aspects, while others are purely linguistic and technical.

2. Language and Ideology

2.1. Some Educational Issues

Inclusive writing is expected by many—albeit not by all—to have an ideological impact on gender equality which it endeavors to represent visually. It is also likely to have a positive educational effect on children and teenagers in schools.

The question remains of the oral utterance of graphemic gender-neutral writing. Oral strategies need to be devised, for instance: <les ambassadeur·rice·s> would become orally “les ambassadeurs et les ambassadrices” (with words in alphabetical order, according to the principles presented in Abily et al. 2016).

Let us take a parallel example in Arabic, where “pioneers” translates into *ruwwād* رواد in the masculine form (resulting from the application of the plural pattern *fu‘āl* فُعَاد to the singular *rā'id* رائد), and *rā'idāt* رائدات in the feminine (by adding the suffix *-āt* ات to the singular). The difference between these two forms is related to Semitic morphology, which resorts, in Arabic, to modifications in patterns and in suffixes, and sometimes in both. The result—which is borrowed from a conference in the United Arab Emirates—was the translation of “pioneers of innovation” into: *rā'idāt wa-ruwwād al-ibtikār* رائدات ورواد الابتكار, where two Arabic words translate the English epicene “pioneers”.

Generally speaking, the matter raised by gender-neutral writing is that of the relations between language and things. Some people, for example, will not use the word “cancer” because they are unconsciously in

the grip of the idea that if you say, “So&so has cancer,” then So&so is closer to getting it. Coming back to our question, the epicene masculine appears as representing humanity as essentially made of male people. Eleanor Roosevelt (Gaer, 2009) just after the end of World War II imposed in the UN Charter the term “Human Rights,” instead of “Man’s Rights” (“les droits de l’homme” in French). The underlying idea was, and still is, that humanity is not by essence made of men.

On the other hand, from a purely philosophical and linguistic point of view, gender-neutral writing is a regression in the arbitrary relation between words (or signs) and things (the signs refer to), as if one needed a law in order to enforce linguistic equality between genders, while such an equality could be represented in languages using epicene words.

Nevertheless, considering the many steps which are still required for humanity to accept full equality between genders, such a philosophical or linguistic regression nowadays emerges as a social necessity. The issue is educational: teaching children at school that genders are equal is a question of quite some momentum.

2.2. Linguistic and Graphemic Aspects: Written Utterances That Cannot Be Vocalized

The main question in the linguistics of writing is that graphemic gender-neutral writing occurrences cannot be orally uttered. Let us take two examples of non-oraisable writing in French:

- Next to the Gare du Nord in Paris, a Moroccan gentleman, named Mr Binebine, used to sell, repair, and install taps (<robinet> in French). He had the idea of putting at the front of his store <RoBINeterie BINE> using capital letters to refer to his name and include it graphically.
- In a French university, the name of the “Faculty of Languages” (<Faculté des Langues>) was turned into: <faculté des ■ang■es>, where <langues> is replaced by <anges>, angels. The use of the black square as a meta-glyph is probably inspired by techniques of censorship or of marking unknown glyphs.

These two examples are not directly related to gender-neutral writing, but they recall that written utterances are not directly related to their oral realization. An advertisement for unisex clothes could be seen in France during the first months of 2018 for a trademark, the name of which was represented as LIU·JO,¹ thus featuring the middle dot of graphemic gender-neutral writing which we will present below. The

1. <http://www.liujo.com/fr/>

symbolic meaning of that point here between the names LIU and JO directly refers to the fact that both names can be either masculine or feminine, and that the trademark is “inclusively unisex”.

3. A Formal Model for Graphemic Gender-Neutral Writing Methods

There are many gender-neutral writing methods: use of both genders (“ladies and gentlemen”), use of gender-neutral words (“people,” “person,” “individual,” etc.), use of gender-neutral pronouns, like, in English, the Spivak pronouns: <e>, , <eir>, <eirs>, <emself> (https://en.wikipedia.org/wiki/Spivak_pronoun). In this paper we will deal with gender-neutral writing involving special graphemes and hence not representable in speech: French *écriture inclusive* <administrateur·rice>, German *binnen-I* <KollegInen>, Spanish *arroba* <trabajador@s>, Greek slashed suffixes <φοιτητές/τριες>, etc. We call these methods *graphemic gender-neutral writing methods*.

In this section we propose a formal model of graphemic gender-neutral writing methods which encompasses French, German, Greek, Italian, Portuguese and Spanish approaches.

NOTATION 1. – Let $\langle \rangle$ denote graphemes. Let w be a two-gender singular-number word in French, German (nouns only), Greek, Italian, Portuguese or Spanish, $w_{\sigma,s}$ its singular-number masculine, $w_{\varphi,s}$ its singular-number feminine, $w_{\sigma,p}$ its plural-number masculine and $w_{\varphi,p}$ its plural-number feminine form. We use the following notation:

- let $C(w)$ be the common phonemic prefix of w_{σ} and w_{φ} , i.e., if g_1, g_2, \dots, g_{i_m} are graphemes (or digraphs), g_i representing the phonemes of w_{σ} and $g'_1, g'_2, \dots, g'_{i_f}$ are graphemes (or digraphs) representing the phonemes of w_{φ} , then $C(w) := g_1, g_2, \dots, g_{i_c}$ such that $g_j = g'_j$ for all $1 \leq j \leq i_c$ and $i_c \leq \min(i_m, i_f)$,²
- in some cases, $C(w)$ differs slightly between the masculine and the feminine version, we will denote these by $C_{\sigma}(w)$ and $C_{\varphi}(w)$;
- let φ_{σ} and φ_{φ} be transformations of grapheme chains;
- let SEPG be a special grapheme called separator grapheme;
- let $\langle + \rangle$ denote the grapheme string concatenator;
- let $SS_{\sigma}(w)$ and $SS_{\varphi}(w)$ be the gender-specific suffixes of $w_{\sigma,s}$ and $w_{\varphi,s}$, i.e., $w_{\sigma,s} = C(w_s) + S_{\sigma,s}(w)$ and $w_{\varphi,s} = C(w_s) + S_{\varphi,s}(w)$. Similarly, let $SP_{\sigma}(w)$ and $SP_{\varphi}(w)$ be the gender-specific suffixes of $w_{\sigma,p}$ and $w_{\varphi,p}$.

2. Note that i_c need not be maximal, as in Greek <φοιτητές/τριες> where the <τ>, albeit common to both suffixes <τές> and <τριες>, is not part of $C(w)$.

DEFINITION 1. – *With the notation above, we define the gender-neutral singular-number form $GN_s(w)$ of w , and the gender-neutral plural-number form $GN_p(w)$ of w as:*

$$GN_s(w) := C(w) + \varphi_\sigma(SS_\sigma(w)) + \text{SEPG} + \varphi_\varphi(SS_\varphi(w)),$$

$$GN_p(w) := C(w) + \varphi_\sigma(SP_\sigma(w)) + \text{SEPG} + \varphi_\varphi(SP_\varphi(w)).$$

Here is how this model can be applied to the six languages we are considering, to obtain the singular-number gender-neutral form of a word:

Language	$C(w)$	$SS_\sigma(w)$	$\varphi_\sigma(SS_\sigma(w))$	SEPG	$SS_\varphi(w)$	$\varphi_\varphi(SS_\varphi(w))$	$GN_s(w)$
French	act	eur	eur	·	rice	rice	acteur- <i>rice</i>
German	Student				in	In	StudentIn
Greek	μαθη	τής	τής	/	τρια	τρια	μαθητής/τρια
Italian	ragazz	o			a	@	ragazz@
Portuguese	alem	ão	ão	/	ã	ã	alemão/ã
Spanish	abogad	o			a	@	abogad@

Similarly, here is how the plural-number form is obtained:

Language	$C(w)$	$SP_\sigma(w)$	$\varphi_\sigma(SP_\sigma(w))$	SEPG	$SP_\varphi(w)$	$\varphi_\varphi(SP_\varphi(w))$	$GN_p(w)$
French	act	eurs	eur	·	rices	rice·s	acteur- <i>rice·s</i>
German	Student				innen	Innen	StudentInnen
Greek	μαθητ	ές	ές	/	τριες	τριες	μαθητές/τριες
Italian	ragazz	i			e	@	ragazz@
Portuguese	alem	ães	ães	/	ãs	ãs	alemães/ãs
Spanish	abogad	os			as	@s	abogad@s

3.1. Graphemic Approaches to Gender-Neutral Writing

There are three main approaches to graphemic gender-neutral writing: the Single-Grapheme Replacement method SINGLE (Italian, Spanish), the Marked Feminine Suffix method MARK (German) and the Suffix-Join JOIN method (French, Greek, Portuguese).

3.1.1. SINGLE

In the *Single Grapheme Replacement* method (SINGLE), one or more gender-specific graphemes are replaced by a single, gender-neutral grapheme, which we call *gender replacement grapheme* (REPG) (e.g., in Spanish, <abogad@s> being the gender-neutral form of <abogados> and <abogadas>). The SINGLE method is used in Italian, Spanish and occasionally in Portuguese.

The cognitive load of the SINGLE method depends on the variety of graphemes represented by the replacement grapheme: if it always represents the same pair of graphemes (for the feminine and the masculine version of the word) then its decoding is easier than if it may need to represent various pairs of graphemes and the reader needs to choose among them to rebuild the original word.

3.1.2. MARK

In the *Marked Feminine Form* (MARK) method, the feminine form—and more generally the feminine grammatical role—is used. To denote gender neutrality, the first grapheme of the feminine suffix is marked, most often by case inversion (as in German <StudentInnen>, or <STUDENTiNNEN>, which are the gender-neutral forms of <Studenten> and <Studentinnen>), but possibly also by preceding it by an underscore (the *gender gap*) or by an asterisk (the *gender star*). The MARK method is used in German.

3.1.3. JOIN

In the *Suffix Join* (JOIN) method, the gender-specific suffixes are joined after the stem and are separated by a specific grapheme, called *gender separator grapheme* SEPG (as in French <étudiant·e>, which is the gender-neutral form of <étudiant> and <étudiante> with a middle dot as separator grapheme, or in Greek <véoc/α>, which is the the gender-neutral form of <véoc> and <véα> with a slash as separator grapheme).

The JOIN approach is used in French (*écriture inclusive*), Greek and Portuguese.

The cognitive load of the JOIN method depends mainly on the size of the first suffix which has to be mentally deleted by the reader in order to obtain the version using the second suffix. Therefore, to evaluate JOIN methods for different languages, we introduce the following notion:

DEFINITION 2. – *In a JOIN method, we call backtrack (BT) the length of the first suffix.*

For example, the backtrack of <administrateur·rice> is 3 since the suffix <eur> of length 3 has to be removed in order to obtain the feminine version <administratrice>. Determination of common prefix and backtrack is done separately for the singular and for the plural number. For example, in the singular of the Portuguese word <cantonês>/<cantonesa>, the common prefix is <canton> (BT=2), while its <cantoneses>/<cantonesas> will have a common prefix <cantones> (and therefore again BT=2).

3.2. Gender Symmetry and Asymmetry

DEFINITION 3. – *Let w be a word. We call a graphemic gender-neutral form $GN(w)$ gender symmetric if, grammatically and visually, $GN(w)$ is at equal distance from w_σ and w_φ .*

If $GN(w)$ is grammatically or visually closer to w_φ , we call it φ -privileging.

If it is grammatically or visually closer to w_σ , we call it σ -privileging.

SINGLE methods are mostly symmetric, for example, the Spanish <abogad@s> is *symmetric* because it is equally close to <abogados> and to <abogadas>, where <close> can be either the Levenshtein distance (both strings can be obtained by a single-character substitution) or the visual resemblance of <@> with <o> and <a> (indeed the grapheme <@> has the shape of an <a> contained in a <o>).

As we will see in § 7, the German MARK method is globally φ -privileging since it is basically the feminine form. In the case of <StudentInnen> is visually very close to the feminine <Studentinnen>. The asymmetry is even stronger for words with unlauded feminine form: in these cases $C(w)$ is gender-specific and therefore the choice of using its feminine form brings $GN(w)$ closer to w_φ than to w_σ , e.g., if w_φ is <Ärztin> (with unlauded <Ä>) and w_σ is <Arzt> (no umlaut), then $GN(w)$ is <ÄrztIn>, which is visually closer to <Ärztin> than to <Arzt>.

JOIN methods lack symmetry because an order has to be chosen between masculine and feminine suffix. Indeed, as the linear chaining of graphemes (mostly) reflects their temporal succession, JOIN methods can *never* be symmetric: one of the two suffixes has to be written first and ipso facto becomes privileged. For example, French *écriture inclusive*, when using the $\sigma\varphi$ order, is σ -privileging: <étudiant·e>, <administrateur·rice>.

We discuss the order of suffixes for the French JOIN method in 8.3 and for the Greek JOIN method in 10.1.

4. Hypotheses for Graphemic Gender-Neutral Writing Methods

4.1. Hypotheses for the SINGLE Method

In order to apply the SINGLE method, we need the following hypothesis to be valid:

HYPOTHESIS 1 (Strong SINGLE Hypothesis). – *Both in the singular and in the plural number, the masculine and feminine versions of a given two-gender word differ by a single grapheme.*

To obtain a gender-neutral version of a given word, it suffices to replace that grapheme by a specific gender-neutral and easily identifiable

grapheme, the *replacement grapheme* REPG. We will investigate for each language: (a) the percentage of words (nouns and adjectives) that satisfy the Strong SINGLE Hypothesis, and (b) whether the proposed replacement grapheme is compatible with the hypothesis.

In some cases there are additional differences between the masculine and feminine versions of a word. For example, a vowel may be accented in one case and not in the other, or an additional letter may appear in front of the replacement grapheme in one case and not in the other, or the replacement grapheme may stand for an empty grapheme in one case and not in the other. For these reasons we state a weaker version of the hypothesis:

HYPOTHESIS 2 (Weak SINGLE Hypothesis). – *Both in the singular and in the plural number, the masculine and feminine versions of a given two-gender word differ by a single grapheme (which may be missing or may be preceded by some other grapheme in one of the two genders) and potentially by the presence or absence of an accent on a grapheme of the common prefix.*

This covers cases such as the Italian masculine <arcaici>, the feminine of which is <arcaiche> (a letter <h> is added), the Italian <figli>, the feminine of which is <figlie> (the replacement grapheme <@> in <figli@> stands for an empty grapheme in the masculine version), and the Spanish <mocetón>, the feminine form of which is <mocetona> (without the acute accent).

Gender-neutral words that satisfy the weak SINGLE hypothesis and not the strong one are asymmetric: in the three examples above, <arcaic@> is σ -privileging (since the absence of the <h> brings the gender-neutral form closer to the masculine one) while <arcaich@> is φ -privileging; <figl@> is σ -privileging, while <figli@> is φ -privileging; <mocetón@> is σ -privileging, while <moceton@> is φ -privileging.

4.2. Hypotheses for the MARK Method

In order to apply the MARK method, we need the following hypothesis to be valid:

HYPOTHESIS 3 (Strong MARK Hypothesis). – *The singular feminine form of a word (noun or adjective) is equal to the singular masculine form followed by a suffix, which is the same for all words of the language. The plural feminine form of a word (noun or adjective) is equal to the singular masculine form followed by a suffix, which is the same for all words of the language.*

By marking the first grapheme of the suffix (either by case inversion, or by preceding it by a <_> or a <*> grapheme), we obtain a gender-neutral version of the word. This method relies on the fact that

all nouns of the language share the *same* suffix, otherwise it becomes difficult for the reader to make the connection between marked grapheme and gender-neutral intention. We will investigate whether this is the case for German.

We also state a weaker version of the hypothesis:

HYPOTHESIS 4 (Weak MARK Hypothesis). – *The singular feminine form of a word (noun or adjective) is equal to the singular masculine form (after removing 0, 1 or 2 graphemes and possibly adding an umlaut) followed by a suffix, which is the same for all words of the language. The plural feminine form of a word (noun or adjective) is equal to the singular masculine form (after removing 0, 1 or 2 graphemes and possibly adding an umlaut) followed by a suffix, which is the same for all words of the language.*

This covers cases such as <Beamt<er>, the feminine version of which is not <*Beamt<er>in> but <Beamt<in> (two graphemes have to be removed from <Beamt<er> before adding the <in> suffix) or such as <Jude>, the feminine version of which is umlauted: <Jüdin>.

MARK gender-neutral forms are, by definition, asymmetric since they are visually closer to the feminine form, and hence are ♀-privileging. In the case of umlauted stems, this property is even stronger: <Jüdin> is closer to the feminine form <Jüdin> than the erroneous *<JudIn>, in which the stem has not been umlauted.

4.3. Hypotheses for the JOIN Method

Finally, in order to apply the JOIN method, we need the following hypothesis to be valid:

HYPOTHESIS 5 (Strong JOIN Hypothesis). – *Whether in singular or in plural number, the masculine and feminine forms of a word must have a common nonempty stem, to which a (possibly empty) suffix has to be added in order to obtain the masculine form, and a different suffix has to be added in order to obtain the feminine form.*

We obtain the gender-neutral form by writing the common stem followed by either the masculine or the feminine suffix, then a separator grapheme SEPG and, finally, the other suffix. This hypothesis makes no assumption on the order of suffixes. We call the length of the first suffix *backtrack*.

If we don't require nonemptiness of the common prefix, then any two words can be combined to form a gender-neutral form, even if they have nothing in common, such as <femme·homme> or <fille·garçon>, so the hypothesis is necessarily *true for all words*. Nevertheless, to keep the cognitive load as low as possible, there are cases where we may ignore slight differences in stems. For example, in the case of Greek words, stems may differ only by accent position, as in the masculine <φ<oi>τ<η>τ<η> (accented on the ultima) and the feminine <φ<oi>τ<η>τ<ρι>ας (accented on

the penult); one can disregard the diacritic and write <φοιτητῆ/τριας> instead of the longer <φοιτητῆ/τήτριας>. To cover this case, we state a weaker version of the JOIN hypothesis:

HYPOTHESIS 6 (Weak JOIN Hypothesis). – *Whether in singular or in plural number, the masculine and feminine forms of a word must have a common (modulo accent position) nonempty stem, to which a (possibly empty) suffix has to be added in order to obtain the masculine form, and a different suffix has to be added, in order to obtain the feminine form.*

This makes it possible for the writer to write accent-independent gender-neutral forms <βάτραχος/ίνα> (σφ order) and <βατραχίνα/ος> (φσ order) with backtracks 2 and 3, instead of the absurd *<βάτραχος/ατραχίνα>, *<βατραχίνα/άτραχος>, which would have backtracks 7 and 8 (!!).

Here again, weakness of the hypothesis can increase the asymmetry of the gender-neutral form and cognitive load. This is not the case in our first example <βάτραχος/ίνα> since accents of both forms appear in the gender-neutral form and guide the reader into reconstructing the gender-specific forms. It is, however, strongly the case in the second example <βατραχίνα/ος> since the reader has to reconstruct the masculine form by mentally repositioning the accent to the penult (<βάτραχος>).

Let us now consider graphemic gender-neutral writing methods for various languages: Italian (§ 5), Spanish (§ 6), German (§ 7), French (§ 8), Portuguese (§ 9), and Greek (§ 10).

5. The Italian SINGLE Method

The Italian SINGLE approach consists in replacing the final vowel of Italian nouns and adjectives by a replacement grapheme, which can be <@> or <*>, e.g., <ragazz@ italian@> or <ragazz* italian*>, instead of <ragazzi italiani e ragazze italiane>.

5.1. History

The feminist publication (Not One Less, 2017) uses the <@> sign (called *chiocciola*) as a gender-neutral graphemic replacement of -o/-a (singular) or -i/-e (plural):

In questo Piano abbiamo scelto di svelare la non neutralità del maschile utilizzando non solo il femminile, ma anche la @ per segnalare l'irriducibilità e la molteplicità delle nostre differenze.³

In the 57-page long (ibid.) booklet the gender-neutral grapheme <@> is used 50 times for 23 different words, mostly for the words <tutt@> (“all_{gn}”⁴) and <ognun@> (“nobody_{gn}”). When, on p. 12, an article has is written in gender-neutral form, a slash-based JOIN method is applied instead: <delle/degli altr@> (“of the_{gn} others_{gn}”).

In 2012, in a report financed by the Region of Tuscany and supported by the *Accademia della Crusca*, Cecilia Robustelli (2012) mentions the asterisk <*> as a gender-neutral grapheme, but discourages its use:

L'uso di forme abbreviate attraverso altri espedienti grafici, come per esempio l'inserimento dell'asterisco al posto della desinenza per indicare che si intende sia la forma maschile sia quella femminile, es. *ragazz** anziché *ragazzo/ragazza* o *ragazzo/a*, è da evitare perché può ostacolare la lettura e la comprensione del testo.⁵

Notice that the JOIN slash-based approach <ragazzo/a> is also mentioned.

We have investigated the validity of the strong and weak SINGLE hypotheses for the Italian language. For this we have extracted from the Italian Wiktionary (version of July 20th, 2019) the declension tables of 12,379 Italian nouns, adjectives and participles. Here are the results, according to their compatibility with the two versions of the hypothesis:

Words Compatible with the Strong SINGLE Hypothesis

We have found the following word classes validating the strong SINGLE hypothesis:

3. “In this plan we have chosen to reveal the non-neutrality of the masculine form by using not only the feminine form, but also the @ sign to signal the irreducibility and multiplicity of our differences.”

4. In this text we will mark gender or gender neutrality in English translations as follows: (a) if the word is of feminine or masculine gender in the source language, its translation will carry the subscript “_♀” or “_♂,” resp., e.g., we translate <pronto> by “ready_♂”; (b) if the word is gender-neutral or epicene in the source language, its translations will carry the subscript “_{gn}” or “_{epi}” respectively, e.g., we translate <pront@> by “ready_{gn}”; (c) and if the word is a gender-neutral personal pronoun, to translate it into English we will use Spivak personal pronouns, e.g., we translate <il/elle parle> by “e talks”.

5. “The use of abbreviated forms through other graphic expedients, as for example the insertion of an asterisk instead of the suffix to indicate that both masculine and feminine forms are meant, e.g., *ragazz** instead of *ragazzo/ragazza* or *ragazzo/a*, is to be avoided because it can hinder reading and understanding of the text.”

Class	Suffixes	#	Typical example
1	o/a, i/e	9,138	tutto/tutta, tutti/tutte
2	a/a, i/e	419	ciclista/ciclista, ciclisti/cicliste
3	e/e, i/i	175	inutile/inutile, inutili/inutili
4	o/a, hi/he	157	stucco/stucca, stucchi/stucche
5	e/a, i/e	132	cantoniere/cantoniera, cantonieri/cantoniere
Total		10,021	

These classes cover 81% of the total amount of words. Class 2 is epicene in the singular number, and Class 3 is epicene in both numbers.

Words Compatible with the Weak SINGLE Hypothesis

We found the following word classes validating the weak but not the strong hypothesis:

Class	Suffixes	#	Typical example
6	o/a, i/he	1,515	arcaico/arcaica, arcaici/arcaiche
7	o/a, ∅/e	353	figlio/figlia, figli/figlie
Total		1,868	

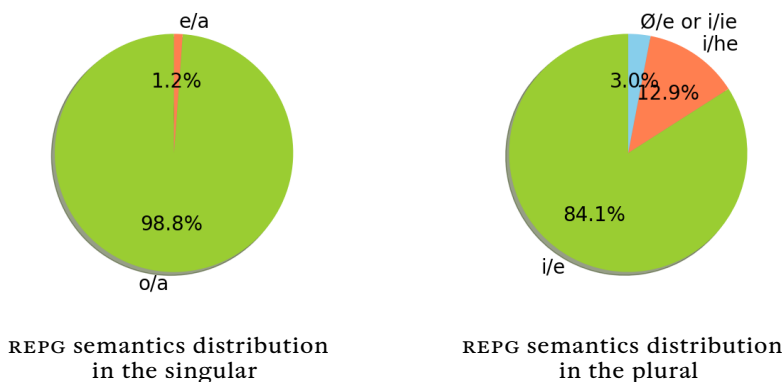
While “strong” Classes 1–5 are totally symmetric (the form <tutt@> can equally well represent <tutti> or <tutte>), this not the case of Classes 6 and 7.

In the plural of Class 6, the writer has to choose between <arcaic@> and <arcaich@>. The former is σ -privileging (since the absence of <h> makes <arcaic@> closer to <arcaici> than to <arcaiche>) and the latter φ -privileging.

In Class 7, the choice is between <figli@> and <figl@>, the first being φ -privileging (since it is more plausible that <@> stands for letter <e> than for a missing letter as in the masculine <figli>) and the second σ -privileging. Ironically, the feminist document (Not One Less, 2017) uses the second form, which is σ -privileging.

Polysemy of the REPG

As we can see in the following diagrams, the semantics of the replacement grapheme are very stable in the singular, since in 98.8% of cases it represents the same pair of vowels o/a. In the plural, only in 84.1% of cases does it represent the pair i/e, while in 12.9% of cases, if we follow the σ -privileging approach, an <h> appears (as in <arcaic@> representing <arcaici> and <arcaiche>).



Forms Incompatible with Both SINGLE Hypotheses

Among the 490 words we found that are incompatible with both versions of the SINGLE hypothesis (that is 4% of the total number of words), let us mention just one class:

Class	Suffixes	#	Typical example
8	ore/rice, ori/rici	422	traduttore/traduttrice, traduttori/traduttrici

Here the suffix differs by significantly more than one letter and hence the SINGLE method cannot be applied. The only possible solution would be to use a JOIN method with a backtrack of 3 letters, to obtain <traduttore/rice>, <traduttori/rici> which, of course, is σ -privileging, since the masculine form comes first. This class represents 86% of the set of incompatible words and 3.4% of the total set of words.

5.1.1. Conclusion

The strong SINGLE hypothesis is valid for 81% of Italian nouns and adjectives, while the weak SINGLE hypothesis is valid for 96% of them. We can reasonably conclude that the SINGLE approach is appropriate for the Italian language.

There is nevertheless a caveat: the replacement grapheme *represents different graphemes in the singular and the plural*, so that the reader must collect information from the context to identify the number of each gender-neutral form, in order to be able to decode it.

6. The Spanish SINGLE Method

The Spanish SINGLE approach involves replacing the vowel of the ultima of Spanish nouns and adjectives by a specific replacement grapheme

which can be <@> (*arropa*), or <*>, or <e>, or <x>: <niñ@s español@s>, or <niñxs españolxs>, or <nin*s español*s>, or <niñes españoles> instead of <niños españoles y niñas españolas>.

6.1. History

In a book called “Sexism and language” (García Meseguer, 1976)⁶, García Meseguer suggests using <e> as replacement grapheme:

Así, cuando uno se dirija a un grupo en una conferencia, en una carta circular, etc., podrá comenzar diciendo *queridos amigos*. *Los trabajadores* podrán escribir en sus pancartas reivindicativas *estamos hartos de ser explotados*. *Los políticos* podrán llamar *compañeros* a sus *partidarios*. *Los progenitores* podrán educar a sus *hijos* más fácilmente en forma no sexista. En los periódicos, los anuncios por palabras solicitarán *una cocinera, una abogada o una secretaria*.⁷

This proposal was followed recently by various politicians: on June 13, 2018, the Argentinian parliamentarian Marcos Cleri started a speech by <Buenas tardes a todes> (“good evening to everybody_{gn}”) and used the <e> replacement grapheme in the entire speech. On June 25 of the same year, the former prime minister of Chile Michelle Bachelet wrote in a tweet: <los miles de chiquilles que hoy estudian con gratuidad en Chile>⁸ In 2018 the National University of La Plata started a television program called *Todes* (“Everyone_{gn}”), and in April 2019 this university organized a Conference on Inclusive Language⁹.

As for other graphemes than <e>, in 2009 the collective publication *Interdicciones, escrituras de la intersexualidad en castellano* (“Interdictions, writings of intersexuality in Spanish,” Cabral 2009) uses the <*> grapheme in several texts.

In 2012 the University of Valencia published a “Guide for an egalitarian language” (Quilis Merín, Albelda Marco, and Josep Cuenca, 2012) where the usage of <@> is discouraged:

6. See also the detailed bibliography in <https://www.sexismoylenguaje.com/polemica-guias-para-un-uso-no-sexis>.

7. “Thus, when you join a group of people in a conference, in a collective letter, etc., you can start by saying *dear friends_{gn}*. *Workers_{gn}* will be able to write in their claim placards “we are *tired_{gn}* of being *exploited_{gn}*”. *Politicians_{gn}* may call their *supporters_{gn}* *companions_{gn}*. *Parents_{gn}* can educate their *children_{gn}* more easily in a non-sexist way. In the newspapers, word ads will request a *cook_{gn}*, a *lawyer_{gn}* or a *secretary_{gn}*.”

8. “The thousands_{gn} of kids_{gn} who study today for free in Chile.”

9. For further information on gender-neutral writing in Argentina, see (Patti, 2018).

Evitar el desdoblamiento abreviado con barras y la arroba (@), a no ser que se trate de nombres propios de organismos, grupos o eventos que la hayan incorporado, o que se emplee como herramienta de diseño publicitario.¹⁰

In 2013, the Argentinian transvestite activist Lohana Berkins promoted the use of graphemes <@> or <x> in the Argentinian newspaper *Página/12* (Berkins, 2013).

In 2016 the government of Chile published a “Guide for a gender-inclusive language,” in which the use of <@> is discouraged:

El signo “@” no es lingüístico, rompe con las reglas gramaticales del idioma y es impronunciable por lo tanto su uso no es recomendable.¹¹
(National Council of Culture and Arts, 2016, p. 6)

6.2. Evaluation

In the following we investigate whether the SINGLE hypotheses are valid for the Spanish language. For this we have extracted from the Spanish Wiktionary (version of July 20th, 2019) the declension tables of 73,473 Spanish nouns and adjectives. Here are the following results according to their compatibility with the two versions of the hypothesis:

Invariant or Gender-Invariant Words

The following forms are either totally invariant, or epicene:

Class	Suffixes	#	Typical example
1	(invariant)	26,704	abrecoches
2	(gender inv.)	2,383	moralista, moralistas
Total		29,837	

These classes cover 40.6% of the total number of words. We will consider that they validate the strong SINGLE hypothesis, since they need no special grapheme in the first place.

10. “Avoid segmentation with slashes or use of the *arroba* (@), with the exception of names of people, organisms, groups or events that have incorporated it, or are using them in advertising design.”

11. “The sign ‘@’ is not linguistic, breaks Spanish language grammar rules and is unpronounceable, therefore its use is not recommended.”

Words Compatible with the Strong SINGLE Hypothesis

The feminine and masculine forms of the following word classes differ only in the final vowel, while adding an <s> for the plural:

Class	Suffixes	#	Typical example
3	o/a, os/as	37,434	abogado/abogada, aboga- dos/abogadas
4	e/es, a/as	13	chilote/chilota, chilotes/chilotas
Total		37,447	

These classes cover 51% of the total number of words.

Words Compatible with the Weak SINGLE Hypothesis

We have found the following word classes, in which the final vowel of the masculine singular form is either missing or different than in the other forms:

Class	Suffixes	#	Typical example
5	∅/a, es/as	3,369	trabajador/trabajadora, tra- bajadores/trabajadoras
6	ón/ona, ones/onas	2,616	mocetón/mocetona, mocetones/mocetonas
7	és/esa, eses/esas	95	montañés/montañesa, montañeses/montañesas
8	án/ana, anes/anas	6	alazán/alazana, alazanes/alazanas
9	ín/ina, ines/inas	6	chapín/chapina, chapines/chapinas
Total		6,092	

These classes cover 8.3% of the total number of words.

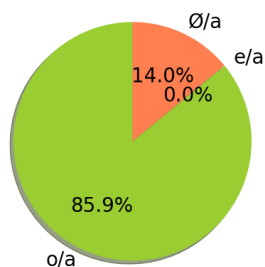
While “strong” Classes 1–4 are totally symmetric (the form <gauch@> represents <gaucho> and <gaucha> equally well and the plural form <gauch@s> represents <gauchos> or <gauchas> equally well), this not the case of Classes 5–9. For example, in Class 5, the form <trabajador@> is ♀-privileging since it assumes the existence of a final vowel, which is only the case for the feminine <trabajadora>. In Class 7, in the singular case, the user has the choice between writing <montañés@> (which is ♂-privileging) or <montañes@> (which is ♀-privileging).

Words Incompatible with Both SINGLE Hypotheses

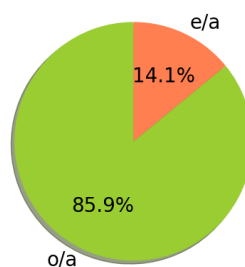
We found 8 nouns with a feminine in *-riz*: <acelerador>/<aceleratriz>, <actor>/<actriz>, <adorador>/<adoratriz>, <director>/<directriz>, <emperador>/<emperatriz>, <formador>/<formatriz>, <generador>/<generatriz>, <tutor>/<tutriz>, <tutriz>, for which only a JOIN method can be used: <actor/riz>, <actores/rizes>, etc., with a backtrack of 2 in the singular and 4 in the plural number.

Polysemy of the REPG

As we can see in the following diagrams, the semantics of the replacement grapheme are relatively stable in both the singular and the plural: in 85.9% of cases it represents the pair of vowels <o>/<a>. Contrary to Italian, for these 85.9% of cases, the replacement grapheme represents the same values <o>/<a> for singular and plural (since, in Spanish, a final <s> morpheme carries the plural information). So, in some sense, the feminine value of the replacement grapheme is always <a>, while the masculine is mostly <o> but can also be empty in the singular or <e> in the plural.



REPG semantics distribution
in the singular



REPG semantics distribution
in the plural

Forms Incompatible with the Use of the <e> Grapheme

Even though the grapheme <e> was historically the first proposed (as early as in 1976, see §6.1), there are many cases in which it cannot be used because the gender-specific suffix already contains an <e>:

- in Class 4, where in both the singular and the plural number, the masculine suffix contains <e>, and the feminine suffix does not;
- in Classes 5–9, where the masculine suffix of the plural number contains <e>, and the feminine suffix does not.

Using an <e> in these cases, which represent 8.3% of the total number of words, results in ambiguity between the gender-neutral and the masculine form: when writing <trabajadores>, is <e> the grapheme of

the masculine plural <trabajadores> or the special gender-neutral replacement grapheme? In the former case <trabajadores> refers to male workers only, while in the latter it refers to workers of both genders.

6.2.1. Conclusion

The strong SINGLE hypothesis is valid for 91.6% of Spanish nouns and adjectives, while the weak SINGLE hypothesis is valid for over 99% of them, when the gender-neutral replacement grapheme is <@>, <*> or <x>. In the case of the <e> grapheme, the weak SINGLE hypothesis cannot be applied, so the total ratio of words compatible with the SINGLE method is only 91.6%.

We conclude that the SINGLE approach is very well adapted to the Spanish language, when one of the gender-neutral replacement graphemes <@>, <*> or <x> is used, but is less efficient when the replacement grapheme <e> is used.

7. The German MARK Method

The German MARK method (called “binnen-I” or “binnenmajuskel”) consists in marking the letter <I> of the feminine suffix of nouns, either by case-inverting it, e.g., <StudentInnen> (where <Studenten> is the masculine plural and <Studentinnen> the feminine plural), or by preceding it by a <_> grapheme or a <*> grapheme, as in <Student_innen> or <Student*innen>. Although it can also be applied to the singular, it is mostly used in the plural, as inclusive of both genders.

When preceded by articles and adjectives, the *feminine* grammatical gender is applied (as in <jede neue KollegIn>, “every_♀ new_♀ colleague_{gn}”), which makes this approach a ♀-privileging one, according to Kotthoff and Nübling (2018, p. 217):

Wegen der Femininkongruenz wird das Femininum (bewusst) privilegiert.¹²

Indeed, the reader’s eye will first recognize the feminine suffix, before (potentially) realizing that letter <I> is in upper case. As (Oestreich, 2009) puts it:

Das Durchschnittsgehirn kennt nämlich keine Binnenmajuskel, also keinen Großbuchstaben inmitten eines Wortes und liest das I als kleinen

12. “The feminine feature is (consciously) privileged, because of gender agreement.”

Buchstaben. Bei PolitikerInnen liest es Politikerinnen – und fragt sich, wo da die Männer bleiben.¹³

As for the singular number, according to Kotthoff and Nübling (2018), adjectives, articles and pronoun dependencies of a gender-neutral noun should use the feminine form: <jede neue KollegIn> (“every_♀ new_♀ colleague_{gn}”). On the other hand, Damm et al. (2014, p. 16) proposes an alternative scheme, where the final grapheme of noun dependencies is also marked: <jedE neuE KollegIn>.

Whether we mark dependencies or not, there is an additional issue which is specific to the singular number, namely *declension*. Indeed, the genitive of the gender-neutral <die ProfessorIn> will be <der ProfessorIn>, which is quite different from the masculine genitive <des Professors>. The result is even more ♀-privileging, and it may be more interesting to write complete words <der Professorin oder des Professors> instead of writing <der ProfessorIn> and have the reader phonetically realize it as <der Professorin oder des Professors>.

7.1. History

According to (Schoenthal, 1998) the MARK method was used for the first time in a self-published 627-page book on pirate radios with instructions on how to build a radio station, published in 1981 (Busch, 1981). In 1983, the MARK method was first used by the Swiss weekly newspaper *Schweizer Wochenzeitung* (WoZ) and the same year by the Berlin newspaper *Berliner Tageszeitung* (taz) (Kotthoff and Nübling, 2018, p. 218).

The MARK method is mentioned in the specific *Duden* on gender issues (Diewald and Steinhauer, 2017, p. 44):

Diese Schreibung ist seit Anfang der 1980er-Jahre belegt und in bestimmten Kontexten sehr gebräuchlich. Allerdings sehen die offiziellen Rechtschreibregeln Binnengroßbuchstaben nicht vor; sie lehnen sie aber auch nicht explizit ab, denn die Binnengroßschreibung ist schlicht gar nicht Gegenstand des amtlichen Regelwerks.¹⁴

In §11 we discuss experimental versions of the MARK method.

13. “The average brain is not aware of *binnen*-letters, i.e., capital letters in the middle of a word, and reads the letter I as a lowercase letter. In the word *PolitikerInnen* (politicians_{gn}) it will read *Politikerinnen* (politicians_♀)—and will wonder why there are no men.”

14. “This form of writing is documented since the early 80s and is very common in specific contexts. However, official spelling rules do not consider internal capital letters; but they do not explicitly prohibit them either, because internal capital writing is simply not an issue for official regulations.”

7.2. Evaluation

To investigate the validity of the MARK hypotheses, we extracted data on 4,561 two-gender nouns from the German Wiktionary.

7.2.1. Strong MARK Hypothesis

We found 4,079 nouns (that is 89.4% of the total number of two-gender nouns) conforming with the strong MARK hypothesis. They can be subdivided into three classes:

Class	#	Typical examples
$GN(w) = C_\sigma(w) + \langle In \rangle$	3,672	Student → StudentIn
$GN(w) = C_\sigma(w) - \text{last letter} + \langle In \rangle$	368	Kollege → KollegIn
$GN(w) = C_\sigma(w) - \text{two last letters} + \langle In \rangle$	39	Beamter → BeamtIn

If i is the number of letters to remove before adding the $\langle in \rangle$ suffix, then $i = 0, 1, 2$ are the only possible cases in our corpus.

7.2.2. Weak MARK Hypothesis

Among the remaining 480 nouns, 136 validate only the weak MARK hypothesis, in the sense that the stem of the feminine noun is unlauded while the stem of the masculine is not. Again we have three classes:

Class	#	Typical examples
$GN(w) = \text{unlauded}(C_\sigma(w)) + \langle In \rangle$	117	Arzt → Ärztin
$GN(w) = \text{unlauded}(C_\sigma(w)) - \text{last letter} + \langle In \rangle$	18	Jude → Jüdin
$GN(w) = \text{unlauded}(C_\sigma(w)) - \text{two last letters} + \langle In \rangle$	1	Tauber → Täubin

7.2.3. Cases Where the MARK Hypothesis Is Invalid

In 26 cases of Wiktionary two-gender nouns, the MARK method cannot be applied because the plural form is epicene: $\langle Angeklagte \rangle$, $\langle Elfe \rangle$, $\langle Linke \rangle$, $\langle Süße \rangle$, etc.

In 193 cases the incompatibility is of a lexical nature: the following gender-related antonymous pairs serve as bases for composite word creation:

Pair	#	Examples
Mann/Frau	54	Ehemann/Ehefrau, Fachmann/Fachfrau, etc.
Vater/Mutter	15	Stiefvater/Stiefmutter, Großvater/Großmutter, etc.
Sohn/Tochter	11	Pflegesohn/Pflegetochter, Enkelsohn/Enkeltochter, etc.
Bruder/Schwester	8	Knastbruder/Knastschwester, Vollbruder/Vollschwester, etc.
Junge/Mädchen	9	Bauernjunge/Bauernmädchen, Zeitungsjunge/Zeitungsmädchen, etc.

In these cases non-graphemic solutions have to be sought.

Finally in about a hundred cases, words are of foreign origin and are feminized according to the rules of their original language: <Coiffeur>/<Coiffeuse>, <Cowboy>/<Cowgirl>, <Filipino>/<Filipina>, <Yogi>/<Yogini>, etc.

7.2.4. Conclusion

The strong MARK hypothesis can be applied to 89.4% of German two-gender nouns and the weak MARK hypothesis to 92.4% of German two-gender nouns. Among the remaining cases, 0.5% are epicene in the plural, 4.2% are of a lexical nature and 2.2% are words of foreign origin.

We conclude that the MARK is relatively well suited for the German language.

8. The French JOIN Method

The French JOIN (called “écriture inclusive”) is a gender-neutral writing method using <·>, <.> or <-> as separator grapheme: <étudiant·e·s>, or <étudiant.e.s>, or <étudiant-e-s> for <étudiants et étudiantes>. It uses the sorting order of gender-specific forms as a criterion for the order of suffixes (cf. §8.3).

8.1. History

Between October 2017 and March 2018, there was an animated debate in the French media concerning gender-neutral writing. The spark that ignited the debate (Manesse and Siouffi, 2019, p. 7) was the publication, on September 22, 2017, in the right-wing daily newspaper *Le Figaro* of the following sentence, taken from a 3rd grade school book (Le Callenec, 2017):

Grâce aux agriculteurs-rices, aux artisan·e·s et aux commerçant·e·s, la Gaule était un pays riche.¹⁵

This debate culminated with a statement by the French Academy:

Prenant acte de la diffusion d'une «écriture inclusive» qui prétend s'imposer comme norme, l'Académie française élève à l'unanimité une solennelle mise en garde. La multiplication des marques orthographiques et syntaxiques qu'elle induit aboutit à une langue désunie, disparate dans son expression, créant une confusion qui confine à l'illisibilité. On voit mal quel est l'objectif poursuivi et comment il pourrait surmonter les obstacles pratiques d'écriture, de lecture – visuelle ou à voix haute – et de prononciation. Cela alourdirait la tâche des pédagogues. Cela compliquerait plus encore celle des lecteurs.

Plus que toute autre institution, l'Académie française est sensible aux évolutions et aux innovations de la langue, puisqu'elle a pour mission de les codifier. En cette occasion, c'est moins en gardienne de la norme qu'en garante de l'avenir qu'elle lance un cri d'alarme: devant cette aberration « inclusive », la langue française se trouve désormais en péril mortel, ce dont notre nation est dès aujourd'hui comptable devant les générations futures.¹⁶ (Académie française, 2017)

The statement about French language being in “mortal danger” seems utterly exaggerated, but may be due to the fact that, until then, the French Academy had been dealing with the acceptability of individual words, and had never to face a meta-technique which applies to tens of thousands of words.

On November 22, 2017, the Prime Minister Édouard Philippe officially prohibited the use of “écriture inclusive” gender-neutral writing in public administration:

[...] je vous invite, en particulier pour les textes destinés à être publiés au *Journal officiel de la République française*, à ne pas faire usage de l'écriture dite

15. “Thanks to farmers_{gn}, craftsmen_{gn} and merchants_{gn}, Gaul was a wealthy country.”

16. “Taking note of the spread of an “inclusive writing” system that claims to become a norm, the French Academy unanimously raises a solemn warning. The multitude of orthographic and syntactic phenomena that it induces leads to a disunited language, disparate in its expression, creating confusion that reaches illegibility. We can hardly identify the goal of this inclusive writing, and we don't see how it could overcome the practical obstacles of writing, reading—visual or aloud—and pronunciation. It would make the task of pedagogues harder. And it would further complicate the task of readers. // More than any other institution, the French Academy is sensitive to developments and innovations in language, since its mission is to codify them. On this occasion, to guarantee the future and to preserve the norm, the French Academy raises an alarm: facing this “inclusive” aberration, the French language is currently in a state of mortal danger, and our nation carries the responsibility of this issue with respect to future generations.”

inclusive, qui désigne les pratiques rédactionnelles et typographiques visant à substituer à l'emploi du masculin, lorsqu'il est utilisé dans un sens générique, une graphie faisant ressortir l'existence d'une forme féminine. Outre le respect du formalisme propre aux actes de nature juridique, les administrations relevant de l'État doivent se conformer aux règles grammaticales et syntaxiques, notamment pour des raisons d'intelligibilité et de clarté de la norme.¹⁷ (Philippe, 2017)

The Prime Minister's reaction to graphemic gender-neutral writing is surprising when we consider the fact that since 2013 his office has been supervising a governmental consulting instance, the *Haut Conseil à l'égalité entre les femmes et les hommes* ("High Council for Equality between Women and Men"), that published in 2016 a guide for gender-neutral writing, including specifications for the French JOIN method.

The appendix of this document (Abily et al., 2016, p. 59–61) contains 96 examples of JOIN gender-neutral forms, using the period <.> as separator grapheme, in singular and plural number. Out of these examples, 15 are epicene. Compared to the classification of French nouns and adjectives that we give in §8.4, the non-epicene examples of (ibid.) are distributed as follows (in %):

Class	1	2	3	4	5	6	7	8	9	10
Ratio in our study	53.3	21.3	6.7	4.6	3	2.6	2.4	2.3	1.7	1.2
Frequency in (ibid.)	28	4	0	5	9	1	16	8	3	3
Ratio in (ibid.)	34.5	4.9	0	6.2	9.4	1.2	19.8	9.9	3.7	3.7

As we see in this table, Class 2 (words with \emptyset /ne, s/nés pattern, such as <doyen-ne-s>) is underrepresented in Abily et al. (ibid.) and Class 3 (pattern \emptyset /e/ \emptyset /es, as in <acquis-e-s>) is completely absent. The absence of Class 3 may be due to the fact that there is a problem in its representation: in other classes, the semantics of a double-separator expression A·B·C are obtained as follows:

1. to obtain the masculine singular form, read A;
2. to obtain the feminine singular form, read AB;
3. to obtain the masculine plural form, read AC;
4. to obtain the feminine plural form, read ABC,

17. "I invite you, especially for texts intended to be published in the *Official Journal of the French Republic*, not to make use of so-called inclusive writing, i.e., the editorial and typographical practices aiming at substituting for the use of the masculine gender, when used in a generic sense, a spelling revealing the feminine-gender form. In addition to respecting the specific editorial rules of legal texts, state administration must comply with grammatical and syntactic rules, in particular for reasons of intelligibility and clarity of the norm."

but in the case of Class 3, Rule 3 is not satisfied: taking the first and third block, we get *<acquiss> instead of <acquis>.

On the other hand, Classes 5 and 7–10 are overrepresented in (Abily et al., 2016), probably because they have the longest suffixes.

For Class 6, only a single example is given (<nombreux·ses>), which is actually a special case since it exists solely in the plural. The absence of more Class 6 examples may be due to the fact that it is the only case where the feminine form of the word is alphabetically sorted before the masculine one (e.g., <peureuse> < <peureux>), and the authors would rather avoid entering into details about this fact. In any case, the example <nombreux·ses> which is given in (ibid.) is wrong: according to the rules explained on p. 27 of the same document, the correct form should be <nombreuses·eux> (the feminine suffix before the masculine one).

Besides a small inconsistency (<sportif·ve> instead of <sportif·ive>), the (ibid.) appendix contains an important mistake: the plural forms of all examples in Class 5 end with -<al·e·s>, implying that the masculine plural should be -<als>, which is absurd: for example, in the case of <principal·e>, the masculine plural is <principaux> and the feminine plural <principales>, therefore the gender-neutral form should be <principales·aux>, instead of *<principal·e·s>. Our hypothesis that this is a mistake is corroborated by the fact that on p. 27 of the same document, the correct version <territoriales·aux> appears as an example.

An interesting case (and a class per se) is the one of word <tout·e>, having the pattern t/te, s/tes. If we follow the rule that the plural of the gender-neutral form is obtained by adding <·s> to the singular gender-neutral form, then the gender-neutral plural should be <tout·e·s>. But this contradicts the Rule 3 given above: using this form, the plural masculine form would be *<touts> instead of <tous>. Therefore Abily et al. (ibid.) recommend the plural gender-neutral form <tou·te·s>, which is suboptimal because it does not have the same stem as the singular form <tout·e>. In Fig. 2 the reader can see graffiti originating from the French spring 2016 student demonstrations; the author of the graffiti was probably unsure about the right spelling of the plural gender-neutral version of <tout·e>: unable to choose between <TOU·TE·S> and <TOUT·E·S>, e merged the two forms and ended up with a form with three (!) separator graphemes.

In 2017, a private communication agency published an additional document containing specifications (Haddad and Baric, 2017), this time using the middle dot <·> as separator grapheme. This document provides the following amendments to Abily et al. (2016):

1. the separator grapheme is a middle dot <·> instead of a period <.> (see also §8.2);
2. the error of Abily et al. (ibid.) concerning the plural of words in Class 5 has been partly corrected: the suffixes are correctly written but their order is still wrong (e.g., the erroneous <local·e·s> of Abily



FIGURE 2. Gender-neutral graffiti <TOU·T·E·S CONTRE LA LOI TRAVAIL!> (“ALL_{gn} AGAINST THE *LOI TRAVAIL!*”), picture taken in Grenoble, in September 2018

et al. (ibid.) has become <locaux·ales>, even though the correct form should be <locales·aux>, cf. §8.3);

3. some additional examples from Class 1 are added, raising the total number of examples of nouns and adjectives to 97;
4. a spelling error is introduced: *<administratr·if> instead of <administrat·if>;
5. the JOIN method is extended to entire words instead of merely suffixes, by writing, e.g., <femme·homme> (“man/woman”). In one case a blank space is even included in the second part of the gender-neutral expression: <du·de la>, where the cognitive load is increased since the reader has to realize that the second part of the gender-neutral expression is not simply <de> but also includes the blank space < > and the following word <la>.

In the period 2018–19 several books on gender-neutral writing in French language have been published, including Manesse and Siouffi (2019), a collective linguistic study of the topic, with information about gender-neutral language issues in English, German, Arabic and Korean.

8.2. Choice of the Middle Dot as Separator Grapheme

Haddad and Baric (2017) justify the choice of the middle dot as follows:

Le point milieu permet d’affirmer sa fonction singulière d’un point de vue sémiotique et par là d’investir « frontalement » l’enjeu discursif et social de l’égalité femmes·hommes.¹⁸ (ibid., p. 9)

What Haddad and Baric (ibid.) probably mean is that they have chosen the middle dot as an unused—and hence totally new in the French—

18. “The middle dot semiotically asserts its specific function and allows a “frontal” investiture of the discursive and social wager of gender equality.”

speaking world—typographical sign, so that it can endow the separator grapheme function as its unique *raison d'être*. This is fundamentally different than the grapheme choices in other methods, such as <I> for German MARK, <@>, <*> and <x> for Spanish SINGLE or the slash </> for Greek JOIN, which are all widely used typographical signs, with a multitude of functions.

Indeed, the middle dot (Unicode 0x00B7 MIDDLE DOT) is used mostly in mathematics (for the multiplication operation, binary operations in algebraic structures, etc.), in Catalan (to separate the two <l> when geminated: <l·l>), in Greek (functioning as a semicolon), in Georgian (functioning as a comma) and in Chinese (as a division marker between transliterated foreign words), but has never been used in French¹⁹. Nevertheless—and this makes it a good choice for a new character to introduce into the French writing system—the middle dot needs no specialized equipment to be inserted into documents: being available in Western-Europe MacRoman and Windows encodings from the beginning, it can be obtained on French MacOS X and Windows keyboard layouts by simple keystroke combinations.

8.3. Order of Suffixes in the French JOIN Method

The order of suffixes in JOIN methods determines whether a given gender-neutral form is σ -privileging or φ -privileging. (Abily et al., 2016) chose to apply the following rule: to *lexicographically compare the masculine and the feminine form* and to use that order for suffixes, e.g., <étudiant> lexicographically sorts after <étudiante>, therefore the gender-neutral form is <étudiant·e> ($\sigma\varphi$ order); <territoriaux> comes after <territoriales>, therefore the gender-neutral form is <territoriales·aux> ($\varphi\sigma$ order).

The hitch is that when a noun has dependencies (articles, pronouns, adjectives), according to agreement rules, all dependencies must keep the same suffix order as the noun, e.g., <les agent·e·s territoriaux·ales>, where the noun <agent·e·s> follows the $\sigma\varphi$ suffix order and therefore the adjective <territoriaux·ales> must follow the same order.

When the noun is epicene, then the adjective is used for suffix-order determination, e.g., in <les fonctionnaires territoriales·aux>, <fonctionnaires> is the noun and therefore the adjective <territoriales·aux> follows its natural $\varphi\sigma$ suffix order.

Abily et al. (ibid.) do not consider the situation when there are many adjective dependencies of the same epicene noun: is it the closest one that determines suffix order for all the others? the longest one? the one on the left or the one on the right? As a puzzle we can consider the

19. It is used though in some French dialects: Occitan, Franco-Provençal and Gallo.

noun with two adjectives <charmant·e fonctionnaire territorial·e>, what would be its plural number? There are two possibilities:

<charmant·e·s fonctionnaires territoriaux·ales>,
<charmante·ant·s fonctionnaires territoriales·aux>,

depending on whether the suffix order is given by the first or by the second adjective. In the first case, the plural backtrack value is equal to 1 for the first adjective and to 3 for the second, in the second case plural backtrack values are 2 and 4, respectively.

8.4. Evaluation of the French JOIN Method

To investigate the validity of the JOIN hypotheses for the French language, we extracted data on 52,271 non-epicene two-gender nouns and adjectives from the French Wiktionary (version of August 1st, 2019), which we divided into 16 classes. We calculated backtrack separately for the singular and for the plural number (values separated by a comma in the BT column):

Class	Suffixes	#	BT	Typical example
1	∅/e, s/es	27,852	0,1	étudiant·e, étudiant·e·s
2	∅/ne, s/nes	11,132	0,1	doyen·ne, doyen·ne·s
3	∅/e, ∅/es	3,521	0,0	acquis·e, acquis·e·s
4	eur/euse, eurs/euses	2,397	3,4	contrôleur·euse, contrôleur·euse·s
5	al/ale, aux/ales	1,583	0,4	principal·e, principales·aux
6	x/se, x/ses	1,367	4,5	peureuse·eux, peureuses·eux
7	eur/rice, eurs/rices	1,278	3,4	directeur·rice, directeur·rice·s
8	er/ère, ers/ères	1,188	2,3	premier·ère, premier·ère·s
9	if/ive, ifs/ives	895	2,3	attentif·ive, attentif·ive·s
10	∅/le, s/les	621	0,1	actuel·le, actuel·le·s
11	∅/te, s/tes	327	0,1	marmot·te, marmot·te·s
12	eau/elle, eaux/elles	41	3,4	beau·elle, beaux·elles
13	∅/que, s/ques	24	0,1	cyprianenc·que, cyprianenc·que·s
14	c/que, s/ques	19	1,2	opoulenc·que, opoulenc·que·s
15	et/ête, ets/êtes	16	2,3	complet·ête, complet·ête·s
16	∅/se, ∅/ses	10	0,0	bas·se, bas·se·s
Total		52,271	0.82	

As can be seen in the table, only Classes 6 (in both numbers) and 5 (in the plural), use the $\varphi\sigma$ order of suffixes: they correspond to merely 5.6% of the total number of words. The value of 0.82 is the weighted average of backtrack values.

8.5. Conclusion

The (strong) JOIN hypothesis has been formulated in such a way that it is valid for all French words, at the expense of potentially high backtrack values. Nevertheless the cases where the backtrack is high are rare when compared to classes such as 1 and 2, for which backtrack is 0. Therefore we observe that the global average backtrack is quite reasonable (less than one grapheme in average) and we conclude that the JOIN is well adapted to French.

8.6. *Rendez-vous pour amant·e·s égaré·e·s:*

An Innovative Use of the French JOIN Method

By common consensus, the gender-neutral expression <un·e étudiant·e> (“a_{gn} student_{gn}”) is generally used with the semantics “a student of either gender”. However, once the gender of a given person is known, the consensus is to use the gender-specific version: <Alice est une étudiante and Bob est un étudiant> (“Alice is a student_♀ and Bob is a student_♂”).

In his 2019 novel *Rendez-vous pour amant·e·s égaré·e·s* (“Appointment for lost_{gn} lovers_{gn}”) (Abbel, 2019), Éric Abbel uses JOIN gender-neutral writing in an innovative way, namely to *bide* the gender of the two protagonists, called <O> and <U>. The 137-page book contains 248 gender-neutral expressions, mostly articles and pronouns, but also adjectives and nouns, and even a pair of inclusively-combined proper nouns: <Ève·Adam> (p. 31).

It is interesting to note that in 121 cases (almost half of the total number of cases), Abbel disobeys rules of JOIN as stated in Abily et al. (2016): he does not respect the order of <elle·il>, <celles·eux>, <jalouse·loux>, and writes <instituteur·trice> instead of <instituteur·rice>. This shows that even though the *écriture inclusive* method has been adopted by many users of French, the rules of suffix order have not yet reached consensus.

Using the JOIN method, Abbel has avoided gender specification, an otherwise difficult task in French, because of the many agreements between articles, nouns, adjectives and pronouns. This challenge has been raised previously by Garréta (1986), without any graphemic method.

Garreta's achievement is comparable to the notorious Oulipian constraints, cf. Becker (2012, Chap. I.3).

In addition to gender obfuscation, Éric Abbel uses another graphemic method: the name of a persona in the novel is written in the Cyrillic alphabet, and therefore is indecipherable for the average French reader: <Одержимый>, a word meaning "obsessed" in Russian and Ukrainian.

Both strategies (gender-neutral JOIN and Cyrillic alphabet) are purely graphemic since they have no phonetic representation, and Abbel may very well be the first author using them in French literature.

9. Portuguese: JOIN or SINGLE?

Gender-neutral writing in Portuguese seems to be divided between the influence of "sister-language" Spanish, for which the SINGLE method is perfectly well suited, and the use of the JOIN method with the slash separator grapheme, which is recommended by academia and provides better coverage of the Portuguese language.

In her 2006 PhD Thesis on gender identity construction in the Portuguese magazine *VIP* (Avanço, 2006), the Brazilian linguist Karla Avanço systematically uses the JOIN method with the slash </> as separator grapheme. In an interview she gave in 2013 to a feminist blog,²⁰ she says:

Tentei usar uma linguagem inclusiva na minha tese e usei um pouco de tudo, menos o "x" e o "@", porque acho que não cabem nesse tipo de texto. Usei com frequência as duas formas "a/o", separadas por barra ou escrevendo as duas palavras, por exemplo: "as leitoras e os leitores",²¹

where by "x and @" she refers to SINGLE methods using <x> and <@> as replacement graphemes, and by "a/o form" she refers to the JOIN method with the slash as separator grapheme.

In 2009, seven years before the French specification (Abily et al., 2016), the Portuguese governmental commission for citizenship and gender equality *Comissão para a Cidadania e Igualdade de Género* published a guide for gender-neutral language in public administration (Abranches, 2009). This guide contains specifications for a JOIN method for Portuguese using the slash (*barrã*) as separator grapheme. It gives 23 examples of gender-neutral forms, 9 of which are epicene, 7 of Class 1 (cf.

20. <https://blogueirasfeministas.com/2013/08/16/linguagem-inclusiva-de-genero-em-trabalho-academico/>.

21. "I tried to use inclusive language in my thesis and used a little of everything except for the "x" and the "@," because I don't think they are suitable for this type of text. I often used the two "a/o" forms, separated by slash or by writing them entirely, as in "the readers_♀ and the readers_♂."

§9.1), 6 of Class 2 and one of Class 5 (the word <cidadã/o>). Suffix order is not mentioned in the specification, but from the examples given we can infer that the order chosen is the one with the least backtrack:

- <a/o cidadã/o> ($\varphi\sigma$) with backtrack 0 is chosen instead of <o/a cidadão/ã> which would have a backtrack of 2;
- <o/a monitor/a> ($\sigma\varphi$) with backtrack 0 is chosen instead of <a/o monitora/or> which would have a backtrack of 3;
- for examples of Class 1 (as in <a/o médica/o> or <o/a beneficiário/a>), suffix order is random since they have backtrack 0 in both cases.

Despite the existence of this specification, there seems to be no consensus in the Portuguese-speaking world.

In 2012, a short text (Oliveira, Duque, and Weyl, 2012, p. 129–132) contained in a collective work on Women’s Law published in Brasilia is devoted to gender-neutral writing and mentions both SINGLE methods (using <@> or <x> as replacement graphemes) and JOIN methods, as in the following sentence:

Em textos alternativos e informais, é possível utilizar o “x” ou mesmo um símbolo como o arroba (a+o=@) para destacar que a/o autor/a esta atenta/o para a linguagem que utiliza,²²

where the authors mention SINGLE methods but in fact use an JOIN method. In the whole text, the SINGLE method is used 5 times, while the JOIN method is used 12 times.

In 2014, a 114-page long “Manual for the non-sexist use of language” (Souza e Silva et al., 2014) published by the government of the Brazilian state Rio Grande do Sul, mentions no graphemic method whatsoever.

Similarly, in 2019, the Brazilian blogger Thaïs Costa arguments²³ against the use of SINGLE methods with <X> and <@> graphemes, but gives no advice about other graphemic methods to use.

9.1. Evaluation

We have investigated the validity of both SINGLE and JOIN hypotheses for the Portuguese language. To that end we have extracted data on 7,799 nouns and adjectives from the Portuguese Wiktionary (as of August 1st, 2019), out of which we have classified 7,700 words into 48 classes, the most important of which are the following eleven:

22. “In alternative and informal texts, it is possible to use an “x” or even a symbol like the at sign (a + o = @) to highlight the fact that the_{gn} author_{gn} is aware_{gn} of the language e uses.”

23. <https://comunidade.rockcontent.com/linguagem-neutra-de-genero/>

Class	Suffixes	#	BT	Typical example
1	o/a, os/as	6,718	1,2	novo/nova, novos/novas
2	ø/a, es/as	472	0,2	observador/observadora, observadores/observadoras
3	ês/esa, eses/esas	107	2,2	cantonês/cantonesa, cantoneses/cantonesas
4	ão/ona, ões/onas	93	2,3	cinquentão/cinquentona, cinquentões/cinquentonas
5	o/ø, os/s	38	1,2	pagão/pagã, pagãos/pagãs
6	ão/ã, ões/ãs	34	2,3	guardião/guardiã, guardiões/guardiãs
7	ø/ø, s/s	34	0,0	inventariante, inventariantes
8	e/a, es/as	34	1,2	presidente/presidenta, presidentes/presidentas
9	o/ø, es/s	26	1,2	alemão/alemã, alemães/alemãs
10	ø/a, s/as	18	0,1	cru/crua, crus/cruas
11	ão/oa, ões/oas	16	2,3	brolhão/brolhoa, brolhões/brolhoas
Total		7,590	1.49	

These classes cover 97.3% of the total number of words extracted. Notice that Class 7 is epicene.

Strong SINGLE Hypothesis

Classes 1 and 8 are the only ones satisfying the strong SINGLE hypothesis, since, for example, <nov@>, <nov@s> and <president@>, <president@s> are perfectly symmetric. If we add to this the epicene Class 7, we find that 89.7% of the total words in the table satisfy the strong SINGLE hypothesis.

Weak SINGLE Hypothesis

Classes 2, 3, 5, 9 and 10 satisfy the weak SINGLE hypothesis: we can write the forms <observador@>, <observador@s> (♀-privileging), <cantones@>, <cantones@s> (♀-privileging), <pagã@>, <pagã@s> (σ-privileging), <alemã@>, <alemã@s> (σ-privileging) and <cru@>, <cru@s> (♀-privileging).

The method does not work for Class 4 (the -ona suffix being too different from the -ão one), for Class 6 (suffixes -ões and -ãs in the plural) and for Class 11 (suffixes -ão and -oa).

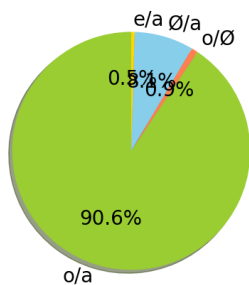
If we add up words satisfying the weak and strong SINGLE hypothesis, we get 98.1% of the words of the table, that is 95.5% of all words extracted from Wiktionary.

We can conclude that the SINGLE method is suitable for Portuguese, even if the large number of irregular forms we found may result in a

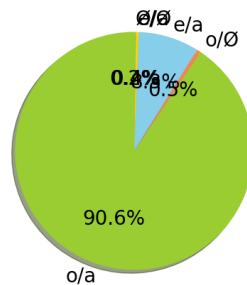
cognitive load for recognizing the grapheme(s) represented by the replacement grapheme.

Polysemy of the REPG

As we can see in the following diagrams, the semantics of the replacement grapheme are quite stable in the singular and in the plural (actually even more than in Spanish): in 90.6% of cases it represents the pair of vowels o/a. The remaining word classes have many suffix pairs (\emptyset/a , o/ \emptyset and e/a for the singular, and e/a, o/ \emptyset , o/ \emptyset , e/ \emptyset and \emptyset/a for the plural), the most important being \emptyset/a for the singular and e/a for the plural. So, like in Spanish, the feminine value of the replacement grapheme is almost always <a>, while the masculine is mostly <o> but can also be empty in the singular or <e> in the plural.



REPG semantics distribution
in the singular



REPG semantics distribution
in the plural

Strong JOIN Hypothesis

The classes represented in the table are all compatible with the JOIN method. Nevertheless the average backtrack we calculated is three times higher than the one for French (§8.4).

9.2. Conclusion

Both methods, SINGLE and JOIN can be used in Portuguese: in the first case, 95.5% of nouns and adjectives of our corpus satisfy the weak SINGLE hypothesis; in the second case, the JOIN hypothesis is satisfied by all words and the average backtrack is higher than the French one, but remains reasonable. The future will show which of the two methods will prevail in Portuguese-language countries.

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικών και Καποδιστριακών
Πανεπιστημίων Αθηνών
— ΙΔΡΥΘΕΝ ΤΟ 1837 —

ΜΕΝΟΥ

1 ΦΟΙΤΗΤΕΣ

2 Ο αριθμός των Ελλήνων και ξένων φοιτητών που εισάγονται κάθε χρόνο στα Πανεπιστημιακά Τμήματα καθορίζεται από το Υπουργείο Παιδείας, αλλά κάθε Τμήμα χωριστά αποφασίζει για τον αριθμό των Ευρωπαϊκών φοιτητών που θα δεχτεί κάθε εξάμηνο στο πλαίσιο των προγραμμάτων κινητικότητας της Ε.Ε. Στην πραγματικότητα, το ΕΚΠΑ ελκύει πολλούς/ές φοιτητές/τριες από ξένες χώρες, με αποτέλεσμα τα τελευταία χρόνια να έχει αυξηθεί σημαντικά ο αριθμός των αλλοδαπών φοιτητών/τριών που φοιτούν στα διάφορα Τμήματα. Ορισμένοι/ες εγγράφονται ως φοιτητές/τριες πλήρους φοίτησης σε προπτυχιακά ή μεταπτυχιακά προγράμματα, ενώ άλλοι συμμετέχουν στα προγράμματα ανταλλαγής φοιτητών/τριών και παρακολουθούν μαθήματα στο ΕΚΠΑ για ένα ή δύο εξάμηνα, έχοντας παράλληλα ενεργή συμμετοχή τόσο στην ακαδημαϊκή ζωή του πανεπιστημίου όσο και στην πλούσια πολιτιστική ζωή μέσα και έξω από το πανεπιστήμιο. Πράγματι, πολλοί/ές ξένοι/ες αλλά και Έλληνες φοιτητές/τριες αισθάνονται προνομιούχοι επειδή ζουν στην ιστορική και κοσμοπολίτικη πόλη της Αθήνας και φοιτούν σε ένα πανεπιστήμιο που έχει διατηρήσει την παράδοση και το κύρος του για 180 χρόνια.

3

4

5

6

7

ΦΟΙΤΗΤΕΣ:
ΠΑΡΟΧΕΣ & ΔΡΑΣΤΗΡΙΟΤΗΤΕΣ
ΗΛΕΚΤΡΟΝΙΚΕΣ ΥΠΗΡΕΣΙΕΣ
ΕΚΠΑΙΔΕΥΤΙΚΑ ΘΕΜΑΤΑ
ΣΥΜΒΟΥΛΕΥΤΙΚΕΣ ΥΠΗΡΕΣΙΕΣ

FIGURE 3. Example of inconsistent use of gender-neutral writing in a page from the official Web site of the University of Athens (<https://www.uoa.gr/foitites/>). The word <φοιτητές> (“students”) appears seven times: twice in masculine form and five times in the gender-neutral syllabic separating form <φοιτητές/τριες>. The sentence containing the 7th occurrence has an agreement error: <πολλοί/ές ξένοι/ες αλλά και *Έλληνες φοιτητές/τριες> (“many_{gn} foreign_{gn} but also *Greek_σ students_{gn}”).

10. The Greek JOIN Method

In Greek, gender-neutral language writing (“μή σεξιστική γραφή,” “non-sexist writing”) uses the JOIN method with the slash </> as separator grapheme. The slash—paradoxically called “κάθετος” (“vertical bar”) even though it is slanted—has a long history in the Greek language since it is widely used for dates, for law numbers, for administrative codes, as well as for contractions (as in <Δ/νσεις Δ/θμιας Έκπ/σης> for <Διευθύνσεις Δευτεροβάθμιας Έκπαιδευσης>), therefore its choice as separator grapheme is a natural one.

Compared with JOIN methods in other languages, the Greek JOIN method is more complex to model and evaluate because of declension: in modern Greek there are four cases (nominative, genitive, accusative, vocative) and we therefore have to take into account four cases, two numbers and two genders, resulting in a total of sixteen forms per word.

10.1. History

One of the first Greek publications on gender-neutral language, the 34-page (Tsokalidou, 1996) contains, already in 1996, a total of ten occurrences of forms: nine nouns and one adjective. The classes of these nouns are distributed as follows (with respect to our classification in §10.2): 8 nouns of Class 10, one noun of Class 31, one adjective of Class 3. The order of suffixes seems to be arbitrary.

In 2006 appears the collective work (Pavlidou, 2006) that systematically uses JOIN in all texts. Among them, Makri-Tsilipakou (2006) uses *exclusively* ♀♂ suffix order, while all other texts use ♂♀ suffix order.

In 2016, the Greek Minister of Internal Affairs released an instruction on the “Insertion of the gender dimension in administrative documents” (Kouvela, 2016), mentioning explicitly gender-neutral forms and the JOIN method:

1. Συστήνεται ή ταυτόχρονη αναφορά σε γυναίκες και άντρες, μέσω της χρήσης και των δύο γραμματικών γενών, όταν το κείμενο αναφέρεται σε μεικτούς πληθυσμούς ή στην περίπτωση που δεν προκύπτει το φύλο.

Αυτό μπορεί να επιτευχθεί είτε με τη χρήση ολόκληρης της λέξης είτε με προσθήκη των καταλήξεων (π.χ. ό/ή διοικούμενος/η ή ό διοικούμενος/διοικούμενη, ό αγρότης/αγρότισσα ή ό/ή αγρότης/ισσα, οί ύποψήφιοι/ες κ.ο.κ.).

[...]

3. Σε περιπτώσεις των ούσιαστικών που ό τύπος του άρσενικού και του θηλυκού ταυτίζονται, συστήνεται τά επίθετα και οί άντωνυμίες να παρατίθενται και στα δύο γένη (π.χ. οί διαθέσιμοι/ες ύπάλληλοι, οί όποιοι/ες...).²⁴

This ministerial instruction institutionalizes JOIN for the Greek language. It refers to a publication of the General Secretary of Gender Equality, the *Guide of Use of non-Sexist Language in Administrative Documents* (Georgallidou et al., 2018) that contains no specifications, but many

24. “1. When a text is referring to mixed populations or when gender is not explicit, we recommend the simultaneous reference to both men and women by the use of both grammatical genders. This can be achieved either by the use of complete words, or by adding suffixes (e.g., the_{gn} governed_{gn} or the_σ governed_σ/governed_φ, the_σ farmer_σ/farmer_φ or the_{gn} farmer_{gn}, the_{epi} candidates_{gn}, and son on).//[...]/3. In the case of epicene nouns, adjectives and pronouns should be written in both genders (e.g., the_{epi} available_{gn} employees_{epi} who_{gn}...).”

examples of gender-neutral forms. Unfortunately it is full of inconsistencies. For example within a few lines, one can find both forms <ὀμιλήτριες/ές> and <ὀμιλήτριες/τές> (p. 19 lines 13 and -4) differing by the absence of grapheme <τ> in the first case. One also finds forms such as <ὀμιλητριῶν/ῶν>, that make no sense since suffixes are not long enough to be different, breaking an implicit rule of the JOIN method which is that suffixes should differ (what is actually meant is <ὀμιλητριῶν/τῶν> with a second suffix of length 3).

A subgroup of the authors of (ibid.) also prepared a similar document for the Observatory of Equality of Cyprus. This document, called *Guide for the Transgression of Linguistic Sexism in the Language of Documents of Public Administration of the Republic of Cyprus* (Gkasouka, Georgallidou, and Foulidou, 2016). Similar to (Georgallidou et al., 2018), it contains a multitude of examples, with, again, a lot of inconsistencies.

As for the order of suffixes, Georgallidou et al. (ibid., p. 42) suggest using the ♀♂ order as much as possible:

Πρόταση: Νὰ χρησιμοποιεῖται συχνὰ ἡ πρόταξη τοῦ θηλυκοῦ γραμματικοῦ γένους. Στόχος εἶναι ἡ ἐπιλογή δήλωσης τοῦ γένους/φύλου νὰ εἶναι ἀνατρεπτικὴ ὡς πρὸς τὸν κυρίαρχο γραμματικὸ κανόνα καὶ νὰ λειτουργήσει ἀφυπνιστικά, ὑποδεικνύοντας τὴ δυνατότητα μιᾶς ἐναλλακτικῆς συντακτικῆς διευθέτησης ὄχι ἀμιγῶς γλωσσικῶν ζητημάτων στὴν ἐκπροσώπηση τῶν φύλων στὸ λόγο.²⁵

Interestingly, the reason invoked is that “this order is *subversive* with respect to the status quo and can contribute to the *awakening* of the reader,” and Gkasouka, Georgallidou, and Foulidou (2016) actually implement this rule throughout the book with the same arguments. Nevertheless, they acknowledge the problem of increased backtrack of the ♀♂ suffix order and add:

Ὁ Ὁδηγὸς εἶναι μὲν ἐξ ὀλοκλήρου γραμμένος μὲ πρόταξη τοῦ θηλυκοῦ τύπου, μὲ σκοπὸ νὰ δείξει ὅτι αὐτὸ ἀποτελεῖ μιὰ πιθανὴ ἐναλλακτικὴ ἐπιλογή τῶν συντακτικῶν/τῶν καὶ γιὰ ἄλλα δημόσια ἔγγραφα, ὡστόσο, τὸ σωστὸ εἶναι πὼς κανένα ἀπὸ τὰ δύο γένη δὲν θὰ ἔπρεπε νὰ δηλώνεται μὲ κατάληξη 3-4 γραμμάτων. Ἐπειδὴ ὅμως αὐτὸ δὲν εἶναι πάντα ἐφικτό, καλὸ εἶναι νὰ ἐναλλάσσονται τὰ γένη ὡς πρὸς τὸ ποῖο προηγεῖται συντακτικὰ καὶ ποῖο ἀκολουθεῖ.²⁶

25. “Recommendation: To use often the feminine-masculine order in suffixes. The objective we pursue is to have the gender/sex declaration to be subversive with respect to the dominant grammatical rule and to contribute in awakening the reader, illustrating the possibility of an alternative syntactic treatment of not entirely linguistic issues in the representation of gender in discourse.”

26. “This Guide is written entirely by using the feminine suffix in the first position, in order to show that this can be a potential alternative author’s choice for other public documents. Nevertheless the right way to proceed is by having the suffix of no gender exceed 3-4 letters. As this is not always possible, one should alternate the order of suffixes.”

In other words, the authors recommend that writers not follow their practice of systematically privileging the feminine suffix, but rather alternate $\text{♀}\sigma$ and $\text{♂}\text{♀}$ suffix orders, not for the sake of gender equality but for practical reasons, as some suffix order may produce very long suffixes.

10.2. Evaluation

In the following we will evaluate the validity of strong and weak JOIN hypotheses for the Greek language and calculate the average backtrack, for both suffix orders.

As it was impossible to extract two-gender nouns from the Greek Wiktionary, we used a different resource: the *Major Greek Dictionary* by Tegopoulos-Fytrakis (Mandala, 1999). From this resource we extracted 15,715 adjectives and 1,033 two-gender nouns. The number of nouns may seem limited, compared for example to those of Wiktionary, but this dictionary does not label words as being both adjectives and nouns, so for example the very common word <φίλος> (“friend”) is labeled only as an adjective.

In the following tables we have classified adjectives and nouns into 37 classes, depending on their decomposition in common prefix and gender-specific suffixes. Here is how to read an entry: in

9	δεξιός	25	ός/ά	οῦ/ᾶς (οὐ/ᾶς)	ό/ά	έ/ά	1.63, 1.38
			οί/ές	ῶν (ών)	οὐς/ές	οί/ές	

we describe Class 9, a typical example of which is the word <δεξιός>. The number of adjectives in this class is 25. The upper part of the split cells contains singular number suffixes: the suffixes of the nominative case are <ός> for the masculine and <ά> for the feminine word, the suffixes of the genitive case are <οῦ> for the masculine and <ᾶς> for the feminine word, etc. In parenthesized italics, we give the suffixes in the monotonic system, whenever these are different from those of the polytonic system (in this case, they are <οὐ> and <ᾶς>). The last column contains the average backtracks for $\text{♂}\text{♀}$ and $\text{♀}\sigma$ suffix orders (sum of backtracks divided by 16, in this case they are 1.63 for $\text{♂}\text{♀}$ and 1.38 for $\text{♀}\sigma$).

10.2.1. Table of Adjectives

Class	Example	#	Nom.	Gen.	Acc.	Voc.	BT
1	φίλος	6,908	ος/η	ου/ης	ο/η	ε/η	1.63, 1.38
2	καλός	5,774	οι/ες ός/ή	ων οῦ/ῆς (οῦ/ῆς)	ους/ες ό/ή	οι/ες έ/ή	1.63, 1.38
3	νέος	1.175	οί/ές ος/α	ῶν (ών) ου/ας	οὺς/ές ο/α	οί/ές ε/α	1.63, 1.38
4	ψυχοπαθής	891	οι/ες ής	ων οῦς (οῦς)	ους/ες ή	οι/ες ής	0, 0
5	ἀγχογόνος	373	εἷς (εἶς) ος	ῶν (ών) ου	εἷς (εἶς) ο	εἷς (εἶς) ε	0, 0
6	ἀγχώδης	311	οι ης	ων ους	ους η	οι η	0, 0
7	γκρινιάρης	196	εις ης/α	ῶν (ών) η/ας	εις η/α	εις η/α	2.63, 1.63
8	αὐτουργός	62	ηδες/ες ός	ηδων/ων οῦ (οῦ)	ηδες/ες ό	ηδες/ες έ	0, 0
9	δεξιός	25	οί ός/ά	ῶν (ών) οῦ/ᾶς (οῦ/ᾶς)	οὺς ό/ά	οί έ/ά	1.63, 1.38
			οί/ές	ῶν (ών)	οὺς/ές	οί/ές	
Total		15,715				Avg	1.46

10.2.2. Table of Nouns

In this table we use the symbol \uparrow whenever the accent of the common prefix of a form is placed one syllable higher than the accent of the common prefix of the other form, e.g., in “ἥς/ \uparrow τρια” the accent of <φοιτητής> is on the ultima of the common prefix, while the accent of <φοιτήτρια> is on the penult of the common prefix. In Class 31 we even have a difference of two syllables between accents, symbolized by the $\uparrow\uparrow$ symbol: the masculine <βάτραχος> is accented on the antepenult of the form, which is also the antepenult of the common prefix, while the feminine form <βατραχίνα> is accented on the penult of the form, which is the ultima of the common prefix.

Class	Example	#	Nom.	Gen.	Acc.	Voc.	BT
10	φοιτητής	363	τής/ \uparrow τρια	τῆ/ \uparrow τριας (τή/ \uparrow τριας)	τή/ \uparrow τρια	τή/ \uparrow τρια	2.63, 4.63

			τές/↑τριες	τῶν/τριῶν (τῶν/τριῶν)	τές/↑τριες	τές/↑τριες	
11	ἐπιβάτης	220	ης/ισσα	η/ισσας	η/ισσα	η/ισσα	1.63, 4.63
			ες/ισσες	ῶν/ισσῶν (ῶν/ισσῶν)	ες/ισσες	ες/ισσες	
12	ἐπιστάτης	101	της/τρια	τη/τριας	τη/τρια	τη/τρια	2.63, 4.63
			τες/τριες	τῶν/τριῶν (τῶν/τριῶν)	τες/τριες	τες/τριες	
13	γλωσσάς	44	άς/ού	ᾶ/οῦς (ᾶ/οῦς)	ά/ού	ά/ού	2.63, 3.63
			ἄδες/ οὔδες (ἄδες/ οὔδες)	ἄδων/ οὔδων	ἄδες/ οὔδες (ἄδες/ οὔδες)	ἄδες/ οὔδες (ἄδες/ οὔδες)	
14	δουλευτής	41	ής/↑ρα	ἦ/↑ρας (ἦ/↑ρας)	ή/↑ρα	ή/↑ρα	1.63, 2.63
			ές/↑ρες	ῶν/ρῶν (ῶν/ρῶν)	ές/↑ρες	ές/↑ρες	
15	καμαριέρης	39	ης/α	η/ας	η/α	η/α	2.63, 1.63
			ηδες/ες	ηδων/ων	ηδες/ες	ηδες/ες	
16	ινδιάνος	37	ος/α	ου/ας	ο/α	ε/α	1.63, 1.38
			οι/ες	ων	ους/ες	οι/ες	
17	καφετζής	28	ής/ού	ἦ/οῦς (ἦ/οῦς)	ή/ού	ή/ού	2.63, 3.63
			ἦδες/ οὔδες (ἦδες/ οὔδες)	ἦδων/ οὔδων	ἦδες/ οὔδες (ἦδες/ οὔδες)	ἦδες/ οὔδες (ἦδες/ οὔδες)	
18	κλέφτης	27	ης/ρα	η/ρας	η/ρα	η/ρα	1.63, 2.63
			ες/ρες	ῶν/ρῶν (ῶν/ρῶν)	ες/ρες	ες/ρες	
19	εὐχέτης	15	ης/ις	η/ιδος	η/ιδα	η/ις	1.63, 3.38
			ες/ιδες	ῶν/↑ιδων (ῶν/↑ιδων)	ες/ιδες	ες/ιδες	
20	ἀρτίστας	13	ας/α	α/ας	α	α	0.38, 0.38
			ες	ῶν (ῶν)	ες	ες	
21	χριστιανός	12	ός/ή	οὔ/ἦς (οὔ/ἦς)	ό/ή	έ/ή	1.5, 1.38
			οί/ές	ῶν (ῶν)	οί/ές	οί/ές	
22	ξάδελφος	11	ος/↑η	ου/↑ης	ο/↑η	ε/↑η	1.75, 1.63
			οι/↑ες	↑ων/ῶν (↑ων/ῶν)	οι/↑ες	οι/↑ες	
23	τουρίστας	10	ας/τρια	α/τριας	α/τρια	α/τρια	1.63, 4.63

			ες/τριες	ῶν/τριῶν (ὠν/τριῶν)	ες/τριες	ες/τριες	
24	πρίγκιπας	10	↑ας/ισσα	↑α/ισσας	↑α/ισσα	↑α/ισσα	1.63, 4.63
			↑ες/ισσες	↑ων/ισσῶν (ων/ισσῶν)	↑ες/ισσες	↑ες/ισσες	
25	ἀθεϊστής	9	ής/↑τρια	ἦ/↑τριας (ῆ/↑τριας)	ή/↑τρια	ή/↑τρια	1.63, 4.63
			ές/↑τριες	ῶν/τριῶν (ὠν/τριῶν)	ές/↑τριες	ές/↑τριες	
26	διδάσκαλος	7	↑ος/ισσα	↑ου/ισσας	↑ο/ισσα	↑ε/ισσα	1.75, 4.63
			↑οι/ισσες	↑ων/ισσῶν (↑ων/ισσῶν)	↑οι/ισσες	↑οι/ισσες	
27	καλλιτέχνης	6	ης/ιδα	η/ιδας	η/ιδα	η/ιδα	1.63, 3.63
			ες/ιδες	ῶν/↑ιδων (ὠν/↑ιδων)	ες/ιδες	ες/ιδες	
28	συμπέθερος	6	↑ος/α	↑ου/ας	↑ο/α	↑ε/α	1.63, 1.38
			↑οι/ες	ων	↑ους/ες	↑οι/ες	
29	μάγος	6	ος/ισσα	ου/ισσας	ο/ισσα	ε/ισσα	1.88, 4.63
			οι/ισσες	↑ων/ισσῶν (↑ων/ισσῶν)	ους/ισσες	οι/ισσες	
30	νονός	5	ός/ά	οῦ/ᾶς (οὔ/ᾶς)	ό/ά	έ/ά	1.5, 1.38
			οί/ές	ῶν (ὠν)	οί/ές	οί/ές	
31	βάτραχος	4	↑↑ος/ίνα	↑↑ου/ίνας	↑↑ο/ίνα	↑↑ε/ίνα	1.88, 3.63
			↑↑οι/ίνες	↑ων/ινῶν (↑ων/ινῶν)	↑↑ους/ίνες	↑↑οι/ίνες	
32	σουρμελής	4	ής/ίδισσα	ἦ/ίδισσας (ῆ/ίδισσας)	ή/ίδισσα	ή/ίδισσα	2.63, 6.63
			ἦδες/ίδισ- σες (ῆδες/ίδισ- σες)	ἦδων/ιδισ- σῶν (ῆδων/ιδισ- σῶν)	ἦδες/ίδισ- σες (ῆδες/ίδισ- σες)	ἦδες/ίδισ- σες (ῆδες/ίδισ- σες)	
33	ἄρραβωνια- στικός	3	ός/ιά	οῦ/ᾶς (οὔ/ᾶς)	ό/ιά	έ/ιά	1.88, 2.63
			οί/ιές	ῶν/ιῶν (ὠν/ιῶν)	οὔς/ιές	οί/ιές	
34	ἄράπης	3	↑ης/ίνα	↑η/ίνας	↑η/ίνα	↑η/ίνα	2.63, 3.63
			↑ηδες/ίνες	↑ηδων/ινων	↑ηδες/ίνες	↑ηδες/ίνες	
35	δούκας	3	ας/ισσα	α/ισσας	α/ισσα	α/ισσα	1.63, 4.63
			ες/ισσες	ῶν/ισσῶν (ὠν/ισσῶν)	ες/ισσες	ες/ισσες	
36	αὐτοκρά- τορας	3	ορας/ειρα	ορα/ειρας	ορα/ειρα	ορα/ειρα	3.63, 4.63
			ορες/ειρες	ὄρων/↑ει- ρων	ορες/ειρες	ορες/ειρες	

37	θεράπων	3	ων/αινα	οντος/αινας	οντα/αινα	ων/αινα	4.13, 4.63	
			οντες/ αινες	όντων/ αινών (όντων/ αινών)	οντες/ αινες	οντες/ αινες		
Total						1,033	Avg	2.19

10.3. Hypotheses

The strong JOIN hypothesis covers all words, but the difference in accent position between the two forms increases the cognitive load, in particular when the accent of the form of the second suffix is not visible, like in Class 28 for the $\sigma\varphi$ <συμπέθεροι/ες>, where the reader's brain has to go through all available accent positions for the feminine form: <συμπέθερες> or <συμπεθέρες>? We have the same problem with $\varphi\sigma$ order in Class 31 <βατραχίνας/ου>: is the masculine form <βάτραχου> or <βατράχου>? The latter issue is due to the phonetic difference between demotic and purified Greek: in purified (and ancient) Greek the antepenult cannot be accented when the ultima is long (genitive <βατράχου>), while this rule is not strict for demotic Greek. Here, the tendency is rather to keep the accent on the same syllable throughout declension (genitive <βάτραχου> for the nominative <βάτραχος>).

For this reason we have introduced a weak version of the JOIN hypothesis, where accent position in the common prefix is not taken into account.

According to the tables above, there is an accent position difference between forms for Classes 10, 14, 19, 22, 24–29, 31 (difference of two syllables), 34 and 36. These classes represent 47% of nouns but only 2.9% of the total set of nouns and adjectives. We can therefore say that Greek validates the strong JOIN hypothesis for 97.1% of words, and the weak hypothesis for all words.

As for the calculation of backtrack, we calculated an average of 1.46 for adjectives and 2.19 for nouns, that is a weighted total average of 1.51. This is higher than the backtrack of French, but very close to the value we calculated for Portuguese.

When applying inverse order $\varphi\sigma$, we obtained a backtrack of 1.23 for adjectives (which is lower than the corresponding value for the $\sigma\varphi$ order), but a high value of 4.01 for nouns. The total average, due to weighting, gives a backtrack value of 1.41, which is lower than the total average $\sigma\varphi$ value.

10.4. Conclusion

If we adopt the weak JOIN hypothesis, then all Greek words are covered by this graphemic gender-neutral writing method. The global backtrack is 1.51 in the $\sigma\varphi$ order of suffixes and 1.41 in the $\varphi\sigma$ order. These are acceptable values, close to those of French and Portuguese. The difference with other languages is that backtrack can sometimes be very high, but this happens only in rare cases, for which it would be better to write the complete forms in the first place.

11. Experimental Methods

We have mentioned in §7.1 that, in German, the underscore grapheme `<_>` has been proposed as an alternative to case inversion: `<Student_innen>` instead of `<StudentInnen>`. This has been criticized by antibinarist and LGBTQI communities as institutionalizing gender binarity by building a fixed binary (masculine/feminine) marking scheme.

To remedy the rigidity of this underscore usage, which is called *static underscore*, they propose the exact opposite: a *dynamic underscore* (“wandernder Unterstrich” in German: a “wandering underscore”) which can be placed anywhere, except at the place where the static underscore would normally be placed. Here is an example (Damm et al., 2014, p. 23):

We lche Mita_rbeiterin will denn i_hre nächste Fortbildung zu anti-diskriminierender Lehre machen? Sie_r soll sich melden. Der Kurs ist bald voll.²⁷

The randomization of the position symbolizes the fact that gender is a continuously changing dynamic process. Besides placing an underscore at random places, the method allows any creative intervention, as in the example above where the feminine pronoun `<Sie>` has been merged with the masculine pronoun `<Er>` to give `<Sie_r>`. In some sense, the idea behind this technique is that it is *not* necessary to produce specific gender-neutral forms, but merely to mark forms in order to show that *gender neutrality is taken into account*. Placing an underscore inside `<Mitarbeiterin>` shows the gender-neutral *intention* of the writer, and this is enough. If orthography is a set of lexical and morphological constraints, there is no need to add more constraints to the existing ones. On the contrary, gender neutrality provides writers with the opportunity to change the rules, so why not change them in a ludic and creative way?

27. “What_{gn} coworker_{gn} would like to have eir training in antidiscriminatory teaching? E should get in touch. The course will soon be full.”

From a linguistic point of view, this method is purely graphemic, as it leaves the phonetic realization of words invariant. Similar to the asterisk that denotes ungrammatical forms in linguistics, it denotes gender-neutrality. But contrarily to other signs that *add* information to a word, this one *invalidates* the existing morphological gender-specific information: the reader is invited to ignore the morphological gender mark and to consider the word as being gender-neutral. In other words, it breaks not only the phonetic mechanism (since it has no phonetic realization) but also the morphological one (since the gender morphemes are deliberately ignored).

The wandering underscore has been used in the title <feministische w_orte> of Hornscheidt (2012), a book on gender studies and gender linguistics. According to Damm et al. (2014, p. 24) (which also uses it in its title *W_ortungen statt Tatenlosigkeit!*), the first use of the wandering underscore in a published text was in Tudor (2010), a text on racism and migrationism that was part of a collective book on racism in Germany.

12. Gender-Neutral Forms as Regular Expressions

Regular expressions were introduced in the seminal paper (Kleene, 1951) where Kleene defines finite automata and shows the equivalence between these finite automata and a class of “events” in “nerve nets” (the way neural networks were called at the time), which he calls “regular events”²⁸. This was five years before Chomsky, in the equally seminal paper (Chomsky, 1956), defined his hierarchy of formal languages, *regular languages* being the simplest ones, and the only ones that can be described by regular expressions. In 1993, regular expressions became standardized as part of the POSIX family of standards (ISO/IEC, 1993).

Given a set called *alphabet* (in our case: the set of graphemes for a given writing system) and an operator called *concatenation* (in our case: grapheme concatenation), we define *formal words* as concatenations of alphabet members (plus a special word called an *empty word*). A *formal language* is simply a set of formal words. Regular expressions serve to describe (potentially infinite) formal languages by writing paradigmatic words using alphabet members as well as a small number of characters (mostly punctuation) with special semantics. For example, in POSIX notation, $(to|ta)\{1,3\}$ represents the formal language of words made out of a single, double or triple concatenation of *to* and/or *ta*, that is the

28. As noted in a footnote in (Kleene, 1951, p. 46), Kleene hesitated to name regular expressions, *prehensible* expressions: “McCulloch and Pitts use a term “prehensible,” introduced rather differently; but since we did not understand their definition, we are hesitant to adopt the term.”

set of formal words {to, ta, toto, tota, tato, tata, tototo, totota, totato, totata, tatoto, tatata, tatata}.

As can be seen in the example (to|ta){1,3}, regular expressions provide symbols for separating alternatives (typically the vertical bar | which is equivalent to a Boolean OR), for grouping (typically pairs of parentheses), and for quantifying.

Graphemic gender-neutral writing methods follow, at least partly, the same agenda: to separate alternatives (i.e., gender-specific suffixes) and to group graphemes (i.e., a single-symbol replacement grapheme representing more than one grapheme). If we consider a gender-neutral form as a regular expression, the formal language it represents is exactly the set of gender-specific forms of the word, e.g., in the case of the German StudentInnen, the formal language would be {Studenten, Studentinnen}.

We can therefore legitimately ask the question: can graphemic gender-neutral expressions be represented by regular expressions? If so, this would mean that a simple rule-based decision process would be sufficient to go back and forth from the gender-neutral form to the set of gender-specific forms. To what extent is this possible? In other words, what is the linguistic background necessary to NLP applications to efficiently identify and decode gender-neutral forms?

Let us consider the three types of graphemic gender-neutral writing.

12.1. SINGLE

Translating a gender-neutral form such as the Spanish <nov@s> into a regular expression is straightforward, provided we know the semantics of the <@> grapheme. As we have seen, for 85.9% of Spanish nouns and adjectives, the value of <@> will be o/a, both in the singular and the plural: the corresponding regular expression will be nov(o|a)s. In cases where the masculine suffix is empty, as in <trabajador@>, we will write trabajador(|a). The NLP application will just have to keep a list of word stems for which <@> takes values other than o/a, and use o/a as a default replacement for the rest.

For Italian the process is more complicated since the same <@> will have different values depending on the number of the word. In a sentence like <quest@ bell@ ragazz@ sono pront@>, it is the verb <sono> (in the plural number) that provides the information that the noun and its dependencies are in plural number, and hence <@> takes values i/e in 84.1% of cases. Its translation into a regular expression would be quest(i|e) bell(i|e) ragazz(i|e) sono pront(i|e).

12.2. MARK

Translating a plural MARK gender-neutral form satisfying the Strong MARK hypothesis is straightforward: <StudentInnen> becomes Student(en|innen). As we have seen in §7.2.1, this works for 89.4% of German two-gender nouns.

The situation is more difficult for words satisfying only the Weak MARK hypothesis: in this case, the pair of parentheses of the regular expression has to encompass also the umlauted vowel, as in <JüdInnen> which translates into J(uden|üdinnen)²⁹. This means that the NLP application will have to detect and process differently umlauted gender-neutral forms³⁰.

The singular number is more problematic, since case has also to be taken into account: <die StudentIn> (nominative) will be translated d(er|ie) Student(|in) while <der StudentIn> (genitive) has to become d(es|er) Student(en|in), since the genitive masculine singular takes its own suffix. The NLP application will have to detect case from the noun's dependencies before translating the gender-neutral expression.

12.3. JOIN

In the case of JOIN the main difficulty for translating into regular expressions will be the backtrack.

Indeed, whenever the backtrack is zero, gender-neutral expressions translate straightforwardly: <étudiant·e> becomes étudiant(|e), <étudiant·e·s> becomes étudiant(|e)s, etc. In French, this is the case for Classes 1–3, 10, 11, 13 and 16, that is 83.2% of the total number of words represented in Table §8.4. For Portuguese this is the case for Classes 2, 7 and 10, that is only 6.9% of words in Table §9.1. And for Greek this *never* happens, as can be observed in Table §10.2.

For the remaining 16.8% of French words, 93.1% of Portuguese word and 100% of Greek words, the NLP application will have to store the backtrack of each word: when knowing that the backtrack of <traducteur·rice> is 3, it will include 3 graphemes before the separator grapheme into the disjunctive pair of parentheses: traduct(eur|rice).

The situation is more complex for Greek. First of all, the value of backtrack depends on the case of the word: <βάρταχος> (nominative)

29. Writing J(u|ü)d(en|innen) seems appealing, but is not correct since the formal language obtained in {Juden, Jüdinnen, Judinnen, Jüden}, where the two last formal words are not German words. This is a superset of the formal language we need, namely {Juden, Jüdinnen}.

30. Except those for which the masculine plural form is also umlauted: the translation of <ÄrztInnen> as Ärzt(e|innen) is straightforward.

has a backtrack value of 2, while <βάρτραχε> (vocative) has a backtrack value of 1. The NLP application will have to detect the word's case and apply the correct backtrack value.

More difficult is the problem of accent position. In classes containing the ↑ symbol in the table of §10.2, the accent is not in the same position of the stem for both genders. In some cases, both accents are visible: <βάρτραχος/ίνα> expands into <βάρτραχος/βατραχίνα> and the NLP application will only need to remove the stem accent when adding the already accented suffix. In other cases, the second-gender (“second” in suffix order) accent is not visible: in <βατραχίνα/ος> with a backtrack of 3 the NLP application knows that the first gender-specific form will be <βατραχίνα> and that the second one is made out of the stem <βατραχ> and the suffix <ος>, but the accent is missing. Once again external resources are needed to detect the accent of this word, knowing its gender, number and case.

12.4. A Possible Future of Gender-Neutral Writing

Regular expressions are part of computing, and computing is more and more pervasive in our daily lives. The emergence of graphemic gender-neutral methods may be related to this trend. Indeed, gender-neutral writing is probably one of the first attempts to add regular expression expressive power into natural language (for many years parentheses have been used for that purpose as in <un(e) enseignant(e)>).

The use of regular expressions in natural language has already been explored in poetry, as in

I need /t(w?o{1,2}) w?r(i|a|ough)te?/.

by American poet Dan Waber, which can be read as “I need to right,” “I need to write,” “I need two rate,” “I need too wrought,” and in many other ways (Waber, 2008, p. 149).

Another poem of his may be more difficult to process by the reader:

/sle[ea]p co-*mes too? (? :me)1,2, but in (? :un|re)fl?its and ?:(? :re)*(? :r?un)? (? :st)?art?s\ . I sta(? :y|ge) up (? :un)*til the (? :wh?e+)+h?ours\ . I c!l(? :a|i)mb (? :<=amb)or) my s?w(? :ay|eigh) (? :in)?to the bed(? :room)?\ . All?one, t?here is (k)?no(? (1)w) goo?d(? :k?night's sle[ea]p)?\ ./

This kind of poetry will probably not become very popular in the near future, but the idea of using regular expression notation to add expressive power to written natural language may nevertheless lead to new graphemic methods. Simply by adding parentheses to JOIN expressions, one can obtain unambiguous and very creative expressions: <administrat(eur·rice)> contains information about backtrack,

and parentheses can be elsewhere than at word end: <(fe·ho)mme> (“women or men”), <p(ossi·roba)ble>, etc.

Of course, inclusion of regular expressions or of similar approaches into written language increases cognitive load for writing and reading, and it is debatable whether this approach is efficient for human communication. But as more and more people learn programming languages and regular expressions are omnipresent in them, their emergence in natural language becomes increasingly probable.

13. Conclusion

Gender-neutral writing is a vast subject and we have merely scratched the surface of a particular subarea, namely graphemic methods. After giving a general model of graphemic gender-neutral writing, we have classified the approaches used in French, German, Greek, Italian, Portuguese and Spanish, into three main methods: SINGLE (the simplest method, where a single symbol replaces two different gender-specific suffixes), MARK (where the feminine form is used, suitably marked to show that gender neutrality is meant) and JOIN (where the masculine and the feminine suffix are both written, separated by a specific grapheme). We have evaluated these methods by their linguistic coverage and the cognitive load required for their use. Finally we have compared graphemic gender-neutral writing methods with regular expressions and have discussed the feasibility of decoding gender-neutral expressions by NLP applications.

References

- Abbel, Éric (2019). *Rendez-vous pour amant·e·s égaré·e·s*. Paris: Espaces insé- cables.
- Abily, Gaëlle et al. (2016). *Pour une communication publique sans stéréotype de sexe. Guide pratique*. Paris: La documentation française.
- Abranches, Graça (2009). *Guia para uma Linguagem Promotora da Igualdade entre Mulheres e Homens na Administração Pública [Guide for an Equality-Promoting Language between Women and Men in Public Administration]*. Lisbon: Comissão para a Cidadania e Igualdade de Género [Governmental Commission for Citizenship and Gender Equality].
- Académie française (2017). “Déclaration de l’Académie française sur l’écriture dite «inclusive»”. <http://www.academie-francaise.fr/actualites/declaration-de-lacademie-francaise-sur-lecriture-dite-inclusive>.

- Avanço, Karla Fernanda Fonseca Corrêa (2006). “Performatividade e constituição das identidades de gênero na revista VIP [Performativity and Constitution of Gender Identities in the Magazine VIP]”. PhD thesis. Universidade Federal de Goiás.
- Becker, Daniel Levin (2012). *Many Subtle Channels: In Praise of Potential Literature*. Harvard: Harvard University Press.
- Berkins, Lohana (2013). “Nosotres y el lenguaje [We and Language]”. <https://perma.cc/E2ER-EUCS>.
- Busch, Christoph (1981). *Was Sie schon immer über Freie Radios wissen wollten, aber nie zu fragen wagten!* Münster: Eigenverlag C. Busch.
- Cabral, Mauro, ed. (2009). *Interdicciones, escrituras de la intersexualidad en castellano [Interdictions, writings of intersexuality in Spanish]*. Córdoba: Anar-rés Editorial.
- Chomsky, Noam (1956). “Three Models for the Description of Language”. In: *Transactions on Information Theory* 2, pp. 113–124.
- Damm, Anna et al. (2014). *Was tun? Sprachhandeln – aber wie? W_Ortungen statt Tatenlosigkeit!* http://feministisch-sprachhandeln.org/wp-content/uploads/2015/04/sprachleitfaden_zweite_auflage.pdf. AG Feministisch Sprachhandeln der Humboldt-Universität zu Berlin.
- Diewald, Gabriele and Anja Steinhauer (2017). *Richtigendern*. Berlin: Duden.
- Gaer, Felice (2009). “Women, International Law and International Institutions: The Case of the United Nations”. In: *Women’s Studies International Forum* 32.1, pp. 60–66.
- García Meseguer, Alvaro (1976). *Sexismo y lenguaje [Sexism and Language]*. Madrid: Cambio 16.
- Garréta, Anne (1986). *Sphinx*. Paris: Grasset.
- Georgallidou, Marianthi [Γεωργαλλίδου, Μαριάνθη] et al. (2018). *Όδηγός Χρήσης μὴ Σεξιστικής Γλώσσας στὰ Διοικητικὰ Έγγραφα [Guide of Use of Non-sexist Language in Administrative Documents]*. Athens: Γενική Γραμματεία Ίσότητας τῶν Φύλων [General Secretary of Gender Equality].
- Gkasouka, Maria [Γκασούκα, Μαρία], Marianthi Georgallidou [Μαριάνθη Γεωργαλλίδου], and Xanthipi Foulidou [Ξανθίπη Φουλίδου] (2016). *Όδηγός Έπέρβασης τοῦ Γλωσσικοῦ Σεξισμού στὴ Γλώσσα τῶν Έγγράφων τῆς Δημόσιας Διοίκησης τῆς Κυπριακῆς Δημοκρατίας [Guide for the Transgression of Linguistic Sexism in the Language of Documents of Public Administration of the Republic of Cyprus]*. Nicosia: Παρατηρητήριο Ίσότητας Κύπρου [Equality Observatory of Cyprus].
- Haddad, Raphaël and Carline Baric (2017). *Manuel d’écriture inclusive*. Paris: Mots-Clés.
- Haralambous, Yannis [Χαραλάμπους, Γιάννης] (2020). *Όδηγός γραφῆς γιὰ τὴν ἑλληνικὴ γλῶσσα [Writing Guide for the Greek Language]*. Athens: Άγρα [Agra].
- Hornscheidt, Antje Lann (2012). *Transdisziplinäre Genderstudien*. Vol. 5: *feministische w_orte: ein lern-, denk- und handlungsbuch zu sprache und diskrim-*

- inierung, gender studies und feministischer linguistik*. Frankfurt a/M: Brandes & Apsel.
- ISO/IEC (1993). “ISO/IEC 9945-2:1993 Information Technology – Portable Operating System Interface (POSIX) – Part 2: Shell and Utilities”.
- Kleene, Stephen C. (1951). *Representation of Events in Nerve Nets and Finite Automata*. Santa Monica, CA: The RAND Corporation.
- Kotthoff, Helga and Damaris Nübling (2018). *Genderlinguistik*. Tübingen: Narr Francke Attempto.
- Kouvela, Fotini [Κουβέλα, Φωτεινή] (2016). “Εγκύκλιος για την Ένταξη της Διάστασης του Φύλου στα Διοικητικά Έγγραφα [Instruction on the Insertion of the Gender Dimension in Administrative Documents]”. Υπουργείο Έσωτερικών και Διοικητικής Ανασυγκρότησης, Γενική Γραμματεία Ισότητας των Φύλων, Αθήνα, 10/3/2016, Α.Π. 652 [Minister of Internal Affairs, Secretary of Gender Equality, Athens, March 10, 2016, Protocol Number 652].
- Le Callennec, Sophie (2017). *Questionner le monde*. Paris: Hatier.
- Makri-Tsilirakou, Marianthi [Μακρή-Τσιλιράκου, Μαριάνθη] (2006). “Συμφωνία/Διαφωνία: Άλληλεγγύη και Αντιπαλότητα στις Συνομιλίες Γυναίκων και Αντρών [Agreement/Disagreement: Solidarity and Adversity in Discussions between Women and Men]”. In: *Γλώσσα - Γένος - Φύλο [Language - Gender - Sex]*. Ed. by Theodossia-Soula Pavlidou [Θεοδοσία-Σούλα Παυλίδου]. Thessaloniki: Ίνστιτούτο Νεοελληνικών Σπουδών [Institute of Modern Greek Studies], pp. 81–117.
- Mandala, Maria [Μανδαλά, Μαρία], ed. (1999). *Μεϊζον Έλληνικό Λεξικό [Major Greek Dictionary]*. Athens: Τεγόπουλος Φυτράκης [Tegopoulos Fu-trakēs].
- Manasse, Danièle and Gilles Siouffi (2019). *Le féminin & le masculin dans la langue*. Paris: ESF sciences humaines.
- National Council of Culture and Arts [Consejo Nacional de la Cultura y las Artes] (2016). “Guía de lenguaje inclusivo de género [Guide of Gender-Inclusive Language]”. <https://www.cultura.gob.cl/wp-content/uploads/2017/01/guia-lenguaje-inclusivo-genero.pdf>.
- Not One Less [Non una di meno] (2017). “Abbiamo un piano. Piano Femminista contro la violenza maschile sulle donne e la violenza di genere [We Have a Plan. Feminist Plan against Male Violence upon Women and Gender Violence]”. https://nonunadimeno.files.wordpress.com/2017/11/abbiamo_un_piano.pdf.
- Oestreich, Heide (2009). “Die Erektion im Text”. *taz*, March 7, 2009, <https://taz.de/Das-Binnen-I-und-die-taz/!5166721/>.
- Oliveira, Rayane Noronha, Ana Paula Duque, and Luana Medeiros Weyl (2012). “Linguagem Inclusiv@: O que é e para que serve?! [Inclusive Language: What Is It and What Is It For]”. In: *O Direito Achado na Rua [Law Found on the Street]*. Vol. 5: *Introdução Crítica ao Direito das Mulheres [Critical Introduction to Women’s Law]*. Ed. by José Geraldo de Sousa Ju-

- nior, Stefanova Apostolova Bistra, and Lívia Gimenes Dias da Fonseca. Brasília: Universidade de Brasília [University of Brasília].
- Patti, Daniela (2018). “El sexismo lingüístico: Un análisis interlingüístico entre español e italiano, con un enfoque particular en el fenómeno del nuevo género neutro en Argentina [Linguistic Sexism: An Interlinguistic Analysis between Spanish and Italian, with a Particular Focus on the Phenomenon of the New Neutral Gender in Argentina]”. Tesi di laurea. Università di Bologna.
- Pavlidou, Theodossia-Soula [Παυλίδου, Θεοδοσία-Σούλα], ed. (2006). *Γλώσσα - Γένος - Φύλο [Language - Gender - Sex]*. 2nd ed. Thessaloniki: Ίνστιτούτο Νεοελληνικῶν Σπουδῶν [Institute of Modern Greek Studies].
- Philippe, Édouard (2017). “Circulaire du 21 novembre 2017 relative aux règles de féminisation et de rédaction des textes publiés au Journal officiel de la République française”. In: *Journal Officiel de la République Française* 272. <https://www.legifrance.gouv.fr/eli/circulaire/2017/11/21/PRMX1732742C/jo/texte>.
- Quilis Merín, Mercedes, Marta Albelda Marco, and Maria Josep Cuenca (2012). *Guía de uso para un lenguaje igualitario (castellano) [Usage Guide for an Egalitarian (Spanish) Language]*. Valencia: Universitat de València.
- Robustelli, Cecilia (2012). “Linee guida per l’uso del *genere* nel linguaggio amministrativo [Guidance for the Use of Gender in Administrative Language]”. https://www.uniss.it/sites/default/files/documentazione/c._robustelli_linee_guida_uso_del_genere_nel_linguaggio_amministrativo.pdf.
- Schoenthal, Gisela (1998). “Von Burschinnen und Azubinnen. Feministische Sprachkritik in den westlichen Bundesländern”. In: *Germanistische Linguistik* 139–140, pp. 9–31.
- Souza e Silva, Evelise de et al. (2014). *Manual para o uso não sexista da linguagem [Manual on the Nonsexist Use of Language]*. Governo do Estado do Rio Grande do Sul [State Government of South Rio Grande].
- Tsokalidou, Petroula [Τσοκαλίδου, Πετρούλα] (1996). *Τὸ φύλο τῆς γλώσσας [The Gender of Language]*. Athens: Σύνδεσμος Ἑλλήνων Ἐπιστημόνων [Society of Greek Scientists].
- Tudor, Alyosxa (2010). “Rassismus und Migratismus: die Relevanz einer kritischen Differenzierung”. In: *Rassismus auf gut Deutsch: ein kritisches Nachschlagewerk zu rassistischen Sprachhandlungen*. Ed. by Adibeli Nduka-Agwu and Antje Lann Hornscheidt. Frankfurt a. M.: Brandes & Apsel, pp. 396–420.
- Waber, Dan (2008). “Regular Expressions as a System of Poetic Notation”. In: *P-Queue* 5, pp. 143–156.

What Are We Calling “Latin Script”?

Name and Reality in the Grammatological Terminology

Wang Yifan

Abstract. The main purpose of this paper is to pose a question regarding the term “script” in the grammatological field, in respect of whether accepted referents match up the definition in previous studies. We follow existing definitions, discuss its nature, and test it against a widely known instance, Latin script. We have concluded that what we call by that name is, in many aspects, not integral to be a single “script” in reality. We thus propose an alternative view on its classification and relevance, with some preliminary analysis on this problem.

1. Background

1.1. The Term “Script”


While there are several known controversial concepts in theoretical grammatology (or maybe graphemics; by this term we refer to the semiotics dedicated to “writing”) in terms of their definitions, most notoriously *grapheme* (Kohrt, 1985; Lockwood, 2001), some key terms, to our knowledge including *script*, have gained general acceptance, rarely been questioned in previous research whenever it has been mentioned.

Script. A collection of letters and other written signs used to represent textual information in one or more writing systems. For example, Russian is written with a subset of the Cyrillic script; Ukrainian is written with a different subset. The Japanese writing system uses several scripts.

(The Unicode Consortium, 2016)

The term *script* is reserved for the graphic form of the units of a writing system. Thus, for example, ‘The Croatian and Serbian writing systems are very similar, but they employ different scripts, Roman and Cyrillic, respectively.’

(Coulmas, 2003, p. 35)

Wang Yifan  0000-0002-4244-3817
Graduate School of Education, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8654, Japan
E-mail: wang-yifan@g.ecc.u-tokyo.ac.jp

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1.
Fluxus Editions, Brest, 2019, p. 91-109. <https://doi.org/10.36824/2018-graf-wang>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

A “script” is just a set of distinct marks conventionally used to represent the written form of one or more languages. [...] Thus we will speak of the “Roman script” or the “Chinese script.” A writing system however is a script used to represent a particular language. [...] We will use the terms “orthography” and “writing system” interchangeably. (Sproat, 2000, p. 25)

script 1 In the study of writing, the graphic form of a writing system. [...] A writing system needs a script for its physical representation [...]. For example, the Roman, Cyrillic, Greek, Russian and runic scripts [...]. (Coulmas, 1996, p. 454)

They all agree that a script is a collection of written symbols that serves *writing systems* or *orthographies*, by which they refer to mechanisms bridging between a particular language and a graphic representation, as shown in a definition below.

Writing System. A set of rules for using one or more scripts to write a particular language. Examples include the American English writing system, the British English writing system, the French writing system, and the Japanese writing system. (The Unicode Consortium, n.d.)

Some authors, as far as we could find, do not explicitly give a definition, but still implicitly assume similar frameworks, as they occasionally make remarks like “Why is Czech written in the Roman alphabet” (Rogers, 2005, p. 182). Similarly in DeFrancis (1989) or Daniels (2009).

Others have slightly different set of terminology, such as:

orthography	conventional spelling of texts, and the principles therefor
writing system	a signary together with an associated orthography
script	in this book, equivalent to <i>writing system</i>

(Daniels and Bright, 1996, pp. xliii–xlv)

Daniels (2018, p. 155), however, has modified the definitions as the following, which he states to be “hopefully uncontroversial”.

- | | |
|--------------------|--|
| (1) orthography | conventional spelling of texts, and the principles therefor |
| (2) script | a particular collection of characters (or signs), used to avoid specifying abjad, alphabet, etc. |
| (3) writing system | a script together with an associated orthography |

It is also worth noting that Sampson (1985; 2015) explicitly treats terms including *script* and *writing system* equivalent. That said, the books still seem to implicitly assume that something is shared among writing systems: “keeping the Roman alphabet [...] but departing from the standard English spelling”.

One interesting thing we would like to point out is that, as one can also see from the above, definitions of *script* are usually accompanied by several instances authors think are notable examples of it, in which Roman (Latin), Cyrillic, and Chinese scripts seem to be most commonly referred to.

1.2. Terminology in This Paper

This paper will use *script*, *orthography*, and *writing system* thereafter in accord with the definition of Daniels (2018) cited in Section 1.1. We also replace *writing system* with *alphabet* in this paper when referring to specific writing system (e.g., *English alphabet*) for convenience sake, as most of the discussion involves alphabetic writing systems. Letters are marked in angular brackets (e.g., <A>) when the glyph or grapheme etc. represented by them are in discussion, without any extra notation, as each meaning should be unambiguous from the context.

Latin script, which will be the central topic of this paper, is also widely known in several aliases, such as *Latin alphabet*, *Roman script*, or *Roman alphabet*. This paper will consistently use the name “Latin script,” regardless of what it is addressed by other authors, as we consider them synonymous in this paper. The names might represent different connotation with historical stages, but the discussion is mostly concerned with recent, if not contemporary materials that no confusion would be expected from the possible contrast.

2. The Nature of a Script

2.1. The Emic Nature of Script

As we have previously seen in Section 1.1, a script is over all understood as “a set of graphical forms used in writing systems.” Now the question is whether each “graphical form” stands for a concrete, objective shape that can be identified across scripts, or a conceptual, subjective item that can only be defined inside a system of script? We believe that the elements in the inventory of a script must be the latter—in other words, *emic* units as introduced by Pike (1954).

The fact can be confirmed by a couple of simple observations. Figure 1 shows a logo once employed as the official logo of NASA (National Aeronautics and Space Administration) of USA. In the picture, the entire graphical shape is intended to be read as “NASA,” but the parts corresponds to <A> are realized without the bar in the middle. It, of course, causes little difficulty being recognized as an instance of <A> nevertheless. This, however, can be a little different when we are writing in Greek script, because the system differentiates <A> (Alpha) and <Λ> (Lambda) exactly by that feature. Greek readers would also recognize the shape itself, but a design that equalizes the two is simply wrong. Thus we can say that a script is not made by picking out needed pieces out of the sea of any imaginable graphical shape. We can accept every kind of shape, may it be untypical, just different system might impose different judgment.



FIGURE 1. An old logo of NASA

There will be another question: is it not that where scripts differ is only how to draw lines between numerous elements, each of which is still a concrete shape which one has encountered? What we see in Figure 2 is a passage intended to be meaningful as English, but each distinct shape corresponding to a letter is made to largely resemble katakana and kanji's skeleton in Japanese writing. Most readers who read English and not Japanese should be able to understand the sentences, although they presumably have never seen such rendering of English alphabet before¹. It may sound surprising, but they are so similar to what usual Japanese characters look like that is almost illegible to those who chiefly read Japanese. From this example we can see that the recognition of each element is not founded on actual instances, but on some essential features the element has in the script.

With these above, we can regard a script as a system that has its own rule set of distinction, and a limited number of elements which are differentiated from each other by internal rules.

2.2. The Writing System–Independence

As in Section 1.1, the common perception is that a script can serve for multiple writing systems, and a writing system may utilize one or more scripts. Meanwhile, we have confirmed in Section 2.1 that a script is by nature an emic system.

The relation between a script and a writing system is comparable to that between a sound system (phonological system) and a language. Basically, the former is a subsystem of the latter. Yet there is a main difference, namely that a script is thought to be shareable between many

1. In case whoever has difficulty reading it: the intended reading is hey guys / can't you read / this sentence? / why can't? 'cause you are japanese (obscure casing).

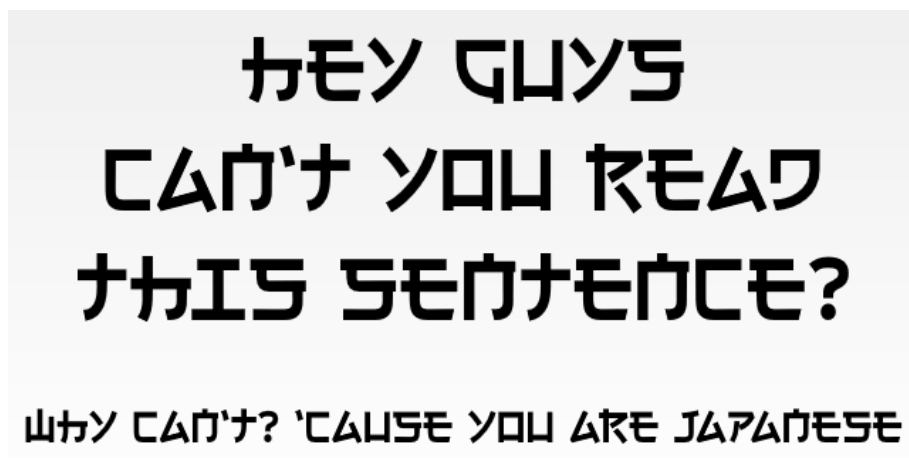


FIGURE 2. Passage in a faux-Japanese² Latin typeface (XYZ4096, 2015)

writing systems. The quality of being independent from writing systems (on the flip side, a writing system can have multiple scripts) is an interesting aspect of script, a unique notion in grammatology. In the world of language, a sound system is usually not considered sharable among multiple spoken languages.

In the same time, it comes with a question: how to determine if scripts used in two writing systems are the same? We could hardly find linguistic literature that discusses the problem, as phonology has been rarely compared across languages. It is an unusual idea to assume many languages share the same sound system or inventory.

2.3. Comparing a Script

Because a script can be shared by writing systems, we need means to compare scripts of different writing systems in order to know whether they are identical. A hint comes from Roy Harris’s works. His theory on writing is, as pointed out by Daniels (1996), admittedly somewhat distant from the “mainstream” writing systems studies cited in Section 1.1, partly because it is tightly combined with the underlying *integrationist*

2. *Note by the Editor.* Alessandrini (1979, p. 44) gives the name *exotypes* to “Latin typefaces that *simulate* non-Latin scripts,” like in the case of this example. (Haralambous, 2007, p. 414).

framework. Despite that, many of his semiotic descriptions are equally meaningful even if we do not presuppose his perspective. He refers to two concepts: *notation* and *script*.

– *notation*:

A notation may serve as a basis for more than one scripts.

(Harris, 2000, p. 92)

A notation may, in principle, serve to articulate any number of different writing systems. Whatever value the figure 5 has [...], it remains recognizable as a member of the series of characters belonging to the notation we call ‘Arabic numerals’.

(Harris, 1995, p. 102)

– *script*:

[T]he typical range of forming and processing activities involved in dealing with letters, numerals, syllabaries, etc. [...] based on the recognition and relative sequencing of the members of an inventory of characters, differentiated [...] by their form.

(ibid., p. 93)

Except that his scope of discussion includes non-glottographic (i.e., which does not translate into oral languages) writing as well, his *notation* and *script* highly resembles *script* and *writing system* in this paper, respectively³. On top of that, Harris (2000, p. 106) provides criteria of a notation.

1. Each member of the set has a specific form which sets it apart from all others in the set.
2. Between any two members there is either a relation of equivalence or a relation of priority. Thus every member has a determinate position with respect to all other members in the set.
3. Membership of the set is closed.

Based on this, we can draw up our criteria to determine when it should be an identical script, summarized as follows:

1. The set has the same *repertoire* of members.
2. The set has the same *boundaries / rules of distinction* among its members.
3. The set has the same set of *internal relationships* among its members.

The criteria has been modified from Harris’s original one by a few points. Firstly, asserting that membership of a script is closed may be too strong, because, while it is probably theory-dependent, some actual writing systems are apparently using an indeterminate number of signs.

3. Beware that our *script* corresponds to Harris’s *notation*, not his *script*.

We lower the hurdle to the level that a consistent mapping of each members will suffice, and merged it into our first and second conditions. Secondly, the second item of Harris’s is too focused on one-dimensional relationship, and we want to augment it to cover any interrelated contrast and/or connection, which is resolved into our second and third conditions. Finally, Harris’s first item becomes a part of our second condition.

3. The “Latin Script” Problem

3.1. Question

Latin script is often cited as the most widespread script that used by majority of the world’s languages (Knight, 1996; SIL International, n.d.). Meanwhile, it is also routinely said that Latin script has been kept monolithic.

[T]hese local forms were always considered to be forms of a single Roman alphabet shared by all western European cultures[...]. If we compare this with the Greek [...] those variants frequently became independent scripts: Coptic, Gothic, Cyrillic, etc. [...] [I]n India a single early script gave rise to a very large number of different scripts. Western Europe, however, maintained a sense of cultural unity which preserved the Roman alphabet intact.

(Rogers, 2005, pp. 175–176)

It declares a belief, that despite the diversity in form and of writing systems adopted in western European languages, they are all founded on an identical set of symbols called Latin script. Is this belief, whereby people call the massive existence in one name “Latin script,” true and valid in the light of its actual function? Could we trust it as a sound grammatological concept? We would like to examine this statement against our criteria described in Section 2.3.

It is to be noted that in subsequent discussions we will only be interested about its solidarity as a script, not other factors related to writing systems. That means we try to isolate what is relevant to comparison of script-level behavior, in the way along the line of previous sections. Topics about correspondence with oral languages and usage of punctuation are out of scope. Some features that characterize a writing system, namely writing direction, digraphs, capitalization, and other rules on combinations of letters in spelling (graphotactics) are excluded because they can be explained as orthographical phenomena. Mentioning differences induced by diacritical elements is also avoided, because we do not have conviction that it does not fall under orthography but script in principle, being merely a vertical version of combination.

3.2. Range of the Latin Script

In order to discuss various properties of Latin script, we must have a definition of the extent it is used. However, there are few exhaustive descriptions available on the extension of the script. Documentations we can temporarily rely on are Wikipedia,⁴ which lists around 150 alphabets counted as Latin script's applications, or ScriptSource (SIL International, n.d.), which lists around 4,500 of them. Here, we will delegate the specification of (commonly acknowledged) Latin-script and non-Latin-script writing systems to those sources for the purpose of discussion.

3.3. Examination

3.3.1. *Repertoire*

ISO basic Latin, recognizing 26 letters, each with two variants—upper and lower cases, could provide us a reasonable starting point of discussion on Latin script repertoire.

ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz

Although the set gains most widespread currency in the Latin-script world, many languages add, drop, or both add and drop base characters, besides the majority of Latin-script orthographies that mandatorily employ diacritics to augment their character sets. This varied coverage, putting other factors aside, presents a difficulty delimiting the extension of Latin script, since as with the classical Sorites paradox, we would be not able to decide how many letters could be added/reduced before an orthography starts/ceases to be a Latin-based writing system. This concern is, unfortunately, something real.

If we are allowed to believe a chart on Wikipedia,⁴ the only two letters that 81 Latin-script orthographies (when accessed back on 2018) agree to have in common are <A> and <I>. What if we adopted it as the minimal requirement of Latin script? Then Belarusian alphabet would be a member by having <A>/<a> and <I>/<i>, contrary to most readers' expectation! What is worse is, at the time of writing of this article, the list has been expanded to include 100 alphabets, the only shared letter of which is <A>.

Can we, on the other hand, define the system by the largest superset? Cherokee script has 5 letters indistinguishable with the basic Latin

4. https://en.wikipedia.org/wiki/List_of_Latin-script_alphabets.

letters (by roman type) in both upper and lower cases, and 17 if limited to the upper case⁵. Even compared to modern Cyrillic inventory, in which less than 10 letters at maximum match that of Latin under the same conditions, it exhibits a great degree of commonality with Latin. Does it mean that Cherokee script is qualified to be incorporated as a variation of Latin? Of course, Cherokee comprises a much larger repertoire of characters that counts over 80, which discounts the fraction of similar letters out of its entirety. But then, what if we had a writing system that used up a large number of additional peculiar letters, which is already partially practiced by some existing African, as well as European alphabets that belong to the Latin script group?

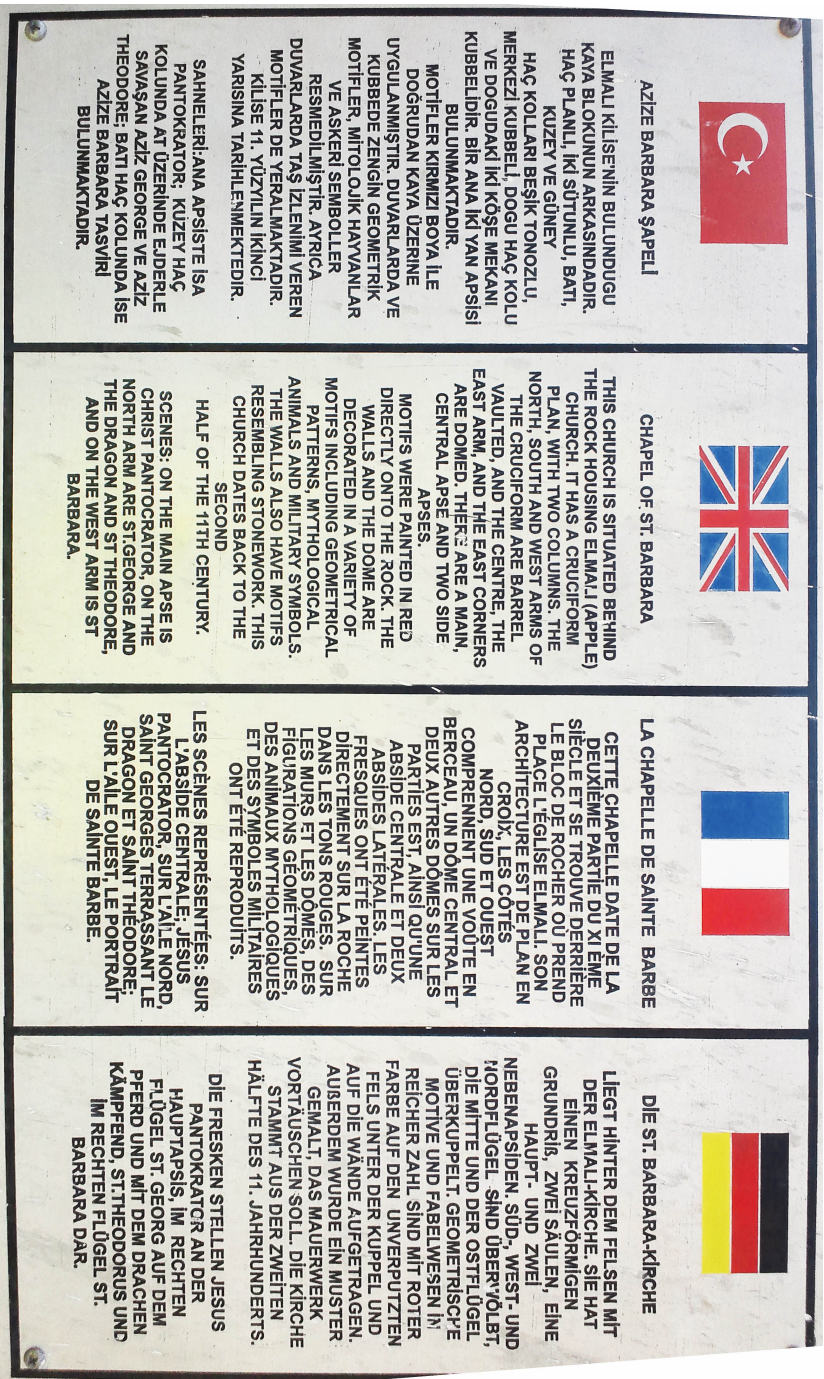
3.3.2. *Distinction*

When we are using a Latin letter in a language, is it the “same” sign we use to write another Latin-based language? Since we have excluded differences in so-called grapheme-phoneme (or we might name it more broadly, “graphic-acoustic”) correspondence, which is naturally idiosyncratic to each orthography, we should reword the question: does a Latin letter share a set of distinctive features throughout Latin-based writing system? This is practically reduceable to a question whether an instance of letter written under an orthography is perceivable as the same letter under another, without misunderstanding.

Perhaps one of the most outstanding, systematic discrepancies among Latin-script writing systems is about the tittle on <I>. Turkish alphabet, with some followers of it, stipulates the dot above <I> to be distinctive, therefore it has two sets of letters <I>/<i> and <İ>/<i> to represent distinct phonemes in the language, where most other Latin-based writing systems ignore the difference, in addition to the fact that <I>/<i> is the standard glyph pair for them in printing. In Turkey, we are never short of examples that (presumably) local literates inadequately stretch their rule to foreign writings. Figure 3 shows a four-language signboard in the all-caps style, which is particularly curious because while the English translation—with no diacritics needed—contains no <İ>, French and German translations—both requiring some diacritics—are printed with all capital <i>’s as <İ>.

The orthographically demanded split of <I> and <İ> has further significance beyond the letters themselves. According to the study of Özer (2016), over the half of subjects (college students attending the calligraphy class) put a tittle above the capital J, although it is an “error,” non-conforming to the prescriptive orthography of Turkish (Figure 6). This

5. The reckoning is based on the number of glyphs rendered identical in the Phoreus Cherokee typeface, which is designed consistently through Latin and Cherokee letters, and reportedly in cooperation with the Cherokee Nation (Jamra, 2015).



AZİZE BARBARA ŞAPELİ

ELMALI KİLİSEİNİN BULUNDUĞU KAYA Bİ OKUNUN ARKASINDADIR. HAÇ PLANLI, İKİ SÜTUNLU, BATI, KUZEY VE GÜNEY HAÇ KOLLARI BEŞİK TONOZLU, MERKEZİ KUBBELİ, DOĞU HAÇ KOLU VE DOĞUDAKİ İKİ KÖŞE MEKANİ KUBBELİDİR. BİR ANA İKİ YAN APSİSİ BULUNMAKTADIR.
MOTİFLER KIRMIZI BOYA İLE DOĞRUDAN KAYA ÜZERİNE UYGULANMIŞTIR. DUVARLARDA VE KUBBEDE ZENGİN GEOMETRİK MOTİFLER, MITOLOJİK HAYVANLAR VE ASKERİ SEMBOLLER RESMEDİLMİŞTİR. AYRICA DUVARLARDA TAŞ İZLENİMİ YEREN MOTİFLER DE YERALMAKTADIR. KİLİSE 11. YÜZYILIN İKİNCİ YARISINA TARİHLENMEKTEDİR.

SAHNELERİ: ANA APSİSTE İSA PANTOKRATOR; KUZEY HAÇ KOLUNDA AT ÜZERİNDE EDERLE SAVŞAN AZİZ GEORGE VE AZİZ THEODORE; BATI HAÇ KOLUNDA İSE AZİZE BARBARA TASYIRI BULUNMAKTADIR.



CHAPEL OF ST. BARBARA

THIS CHURCH IS SITUATED BEHIND THE ROCK HOUSING ELMALI (APPLE) CHURCH. IT HAS A CRUCIFORM PLAN, WITH TWO COLUMNS, THE NORTH, SOUTH AND WEST ARMS OF THE CRUCIFORM ARE BARREL VAULTED, AND THE CENTRE, THE EAST ARM, AND THE EAST CORNERS ARE DOMED. THERE ARE A MAIN, CENTRAL APSE AND TWO SIDE APSES.
MOTIFS WERE PAINTED IN RED DIRECTLY ONTO THE ROCK. THE WALLS AND THE DOME ARE DECORATED IN A VARIETY OF MOTIFS INCLUDING GEOMETRICAL PATTERNS, MYTHOLOGICAL ANIMALS AND MILITARY SYMBOLS. THE WALLS ALSO HAVE MOTIFS RESEMBLING STONEWORK. THIS CHURCH DATES BACK TO THE SECOND HALF OF THE 11TH CENTURY.
SCENES: ON THE MAIN APSE IS CHRIST PANTOCRATOR, ON THE NORTH ARM ARE ST. GEORGE AND THE DRAGON AND ST. THEODORE, AND ON THE WEST ARM IS ST. BARBARA.



LA CHAPELLE DE SAINTE BARBE

CETTE CHAPELLE DATE DE LA DEUXIÈME PARTIE DU XIÈME SIÈCLE ET SE TROUVE DERRIÈRE LE BLOC DE ROCHER OÙ PREND PLACE L'ÉGLISE ELMALI. SON ARCHITECTURE EST DE PLAN EN CROIX, LES CÔTES NORD, SUD ET OUEST COMPRENNENT UNE VOUTE EN BERCEAU, UN DÔME CENTRAL ET DEUX AUTRES DÔMES SUR LES PARTIES EST, AINSI QU'UNE ABSIDE CENTRALE ET DEUX ABSIDES LATÉRALES. LES FRESQUES ONT ÉTÉ PEINTES DIRECTEMENT SUR LA ROCHE DANS LES TONS ROUGES. SUR LES MURS ET LES DÔMES, DES FIGURATIONS GEOMETRIQUES, DES ANIMAUX MYTHOLOGIQUES ET DES SYMBOLES MILITAIRES ONT ÉTÉ REPRODUITS.

LES SCÈNES REPRÉSENTÉES: SUR L'ABSIDE CENTRALE: JÉSUS PANTOCRATOR, SUR L'AILE NORD, SAINT GEORGES TERRASSANT LE DRAGON ET SAINT THEODORE; SUR L'AILE OUEST, LE PORTRAIT DE SAINTE BARBE.



DIE ST. BARBARA-KIRCHE

LIEGT HINTER DEM FELSEN MIT DER ELMALI-KIRCHE. SIE HAT EINEN KREUZFÖRMIGEN GRUNDRIß, ZWEI SÄULEN, EINE HAUPT- UND ZWEI NEBENAPSIDEN. SÜD-, WEST- UND NORDFLÜGEL SIND ÜBERWÖLBT, DIE MITTE UND DER OSTFLÜGEL ÜBERKUPPELT. GEOMETRISCHE MOTIVE UND FABELWESEN IN REICHER ZAHL SIND MIT ROTER FARBE AUF DEN UNVERPUTZTEN FELS UNTER DER KUPPEL UND AUF DIE WÄNDE AUFGETRAGEN. AUßERDEM WURDE EIN MUSTER GEMALT, DAS MAUERWERK VORTAUSCHEN SOLL. DIE KIRCHE STAMMT AUS DER ZWEITEN HÄLFTE DES 11. JAHRHUNDERTS.

DIE FRESKEN STELLEN JESUS PANTOKRATOR AN DER HAUPTAPSIS, IM RECHTEN FLÜGEL, ST. GEORG AUF DEM PFERD UND MIT DEM DRACHEN KÄMPFEND, ST. THEODORUS UND IM RECHTEN FLÜGEL, ST. BARBARA DAR.

FIGURE 3. A signboard in Turkey with dotted I



FIGURE 4. Variations of *ij* (Tubantia – beeld RD, Anton Dommerholt)



FIGURE 5. Variations of *ż* (My another account, 2014)

example clearly shows the result of an analogical induction that what the capital of <j> should look like, when that of <i> is <İ>. We can say that the Turkish system has afforded the conceptualization of titles as a diacritic, unlike other branches of Latin-script alphabets.

Handwritten J's										Doğru Harf	Harflerde Yapılan Yanlışlar	Öğrenci Sayısı
1	2	3	4	5	6	7	8	9	10	Doğru Harf	Noktalı yapanlar	22
11	12	13	14	15	16	17	18	19	20		Altı uzantısını bombeli yapmayanlar	31
21	22	23	24	25	26	27	28	29	30		Harfi sıralamasında yazmayanlar	1
31	32	33	34	35	36	37	38	39	40		Başlangıç bölümünü bombeli yapanlar	2
											Başlangıç dalgasını hiç yapmayanlar	5
										Düzgün olmayanlar	38	

FIGURE 6. Handwritten J's of undergraduate Turks (Özer, 2016)

We can still find examples if we narrow down the scope to Indo-European languages. For example, the <Y>-like glyph is often written in the place of what is usually represented by digraph <IJ> in Dutch (Figure 4), while most of non-Dutch, suppose English, readers would equate it with <Y>. In another case, <Z> and <Ż> are distinct letters in Polish because the latter is a variant of <Ź> (Figure 5), against the conception in some writing systems such as that of English.

These discrepancies signify the difference of distinctive criteria. If an English and a Turkish, an English and a Dutch, or an English and a Polish reader disagree with the identity of a certain glyph, those systems cannot be identical. The situation is comparable to that where the same sample of voice steadily invokes associations with different phonemes for two speakers: they are considered to have different sound systems, which means they speak different languages or dialects.

3.3.3. Ordering

Latin script maintains a certain relatively stable sorting order, which appears to be a hopeful trait to characterize the system if putting aside the status of letters with diacritics. However, according to Comrie (1996), the Lithuanian alphabet disagrees with ISO basic alphabet by putting <Y> between <I> and <J> (because <Y> represents the long vowel of <I>), and so does the Estonian alphabet, with <Z> between <S> and <T> (<Z> being a foreign letter whose sound is akin to that of <S>).

3.3.4. Case

Casing does not account for the uniqueness of Latin script by its own, yet is still possible to be an auxiliary measure. Most variations of Latin script certainly are bicameral, but casing in Saanich alphabet is very



FIGURE 7. Misspelled broken script in Germany (Kobayashi, 2012)

marginal, if not nonexistent. It consists of 38 uppercase glyphs with lowercase <s>, while <s> exclusively marks the third person possessive suffix, which is not exchangeable with the uppercase counterpart (Harvey, 2009).

3.3.5. *Diachrony*

We would like to make some mention of related matters in the diachronic perspective. It is frequently argued that historical glyphs appearing in old documents are also variants of Latin script. Is it true that they are merely allographic to modern glyphs of the script?

Firstly, of course, we have issues in identity of character set, where the classical repertoire of Latin script lacks <I>-<J> and <U>-<V> distinctions alongside an independent <W>, as compared to ISO basic alphabet. But can we still deem that the remaining letters are conceptually unaltered over the course of time?

Akira Kobayashi, a Germany-based typographer, has reported his interesting discovery on broken script (a.k.a. Gothic or Fraktur) misuse (Kobayashi, 2012). Figure 7 shows a sticker intended to be read “Eintracht Frankfurt,” but actually typeset “Eintracht Frantzfurt”. This kind of error suggests whoever in charge of this product understands broken script glyph shape merely by imposing an Antiqua (i.e., contemporary)

H [] GUNZE SANGYO AG ITALERI PAINT NO.		
H[1] 1	ホワイト	1745
H[2] 2	ブラック	1747
H[3] 3	レッド	1503
H[8] 8	シルバー	1546
H[12] 33	フヤ消しブラック	1749
H[18] 28	黒鉄色	1415
H[32] 40	フールドクレー (1)	
H[35] 80	コバルトブルー	2715
H[47] 41	レッドブラウン	1533
H[60] 16	濃緑色	
H[70] 60	RLMグレー02	1591
H[301] 301	グレー-FS36081	
H[305] 305	グレー-FS36118	
H[814] 314	ブルー-FS35622	1731

FIGURE 8. Misspelled Japanese manual (Kimura-mo, 2017)

mental image, and such knowledge does not automatically provide correct discrimination ability of the broken script. Such a situation is, in fact, typically observed when a writer tries to handle non-native writing systems. Figure 8 is a well-known example among Japanese scale model hobbyists, where the imported brand regularly confuses similar-looking characters in Japanese. The cause of such confusion in this case is clearly the unfamiliarity of the writer with Japanese scripts⁶. Even though it has been only seventy years since the ban of broken script in Germany, does not the fact people make similar mistakes imply that the broken script is already a foreign script to the current population? The problem here is practically same as the one we mentioned in Section 3.3.2, and poses a serious question of alleged solidarity of historical Latin script varieties.

3.3.6. Others

Despite all internal differences adduced above, we can observe greater commonalities shared by (at least modern) members of the Latin-script sphere, such as vertical layout including ascender and descender, basic

6. The errors include mistaking し (shi) for レ (re), フ (fu) for つ (tsu), and ワ (wa) for ク (ku).

anatomy of letterforms including stroke and dot, as well as their conjunctions, set of known stylistic variations including italic and boldface. Moreover, if we loosen restriction taking the rough correlation between shape and expected phonetic/phonological value into account, the overall similarity appears more manifest.

So, are these common features altogether sufficient to define Latin script? We consider that it will be also difficult to defend this hypothesis against the notion such as (modern) Cyrillic script, a sibling of Latin script, which already has various features in common, not to mention several homographs with similar phonetic output.

There is another possible argument, namely that even when one acknowledges incomplete agreement of each point stated in previous sections, one can still make up a valid definition by combining the common internal relations above with elements confirmed free of distinctiveness gap, i.e., “particular letters α and β , if exist, must be in the repertoire in this order, and/or $n\%$ of characters must be compatible with a certain set...” The problem with this approach is that it is overly artificial and ad hoc if considering the wild disparity in repertoire, especially without guarantee to be true for future applications of the script.

Shrinking the scope of “Latin script” and regarding most of writing systems virtually as multi-script systems of “Latin” and some idiosyncratic scripts may also be a solution (see Sections 1.1, 2.2), but it ends up in the same problem whether one can distinguish similar scripts by the remaining features.

4. Discussion

After the examination in Section 3.3, we understand that it is hard to justify what has been called the Latin script as a well-defined solid idea. Is there a viable way to encompass the traditional notion of Latin script in its entirety, without letting it be a ship of Theseus? Or do we have to discard the idea from grammarology? We believe the notion equivalent or akin to the current understanding of Latin script still has relevance and importance, just in some other ways.

We think what explains the current situation of Latin script better is such words like *family resemblance* by Wittgenstein (2009) or *prototype* by cognitive linguists (Taylor, 1995). The cognacy inside those writing systems is undeniable, only it is intermediated by mutual similarity between certain single systems, instead of a standard to conform. It forms a vague but continuous concept as much as a rainbow with all of its gradation. After all, the historical truth is that its identity as Latin script has been handed down through repetitive borrowing, adaptation, and/or systematic imitation, rather than consistent rules.

Therefore, we propose to treat the concept Latin script as a genealogical clade, an analogue of family or branch in comparative linguistics. That is, we argue against the view that writing systems described in Section 3.2 shares a common system called Latin script, in favor of one that what they have are multiple “sister scripts” that are related yet still incommensurable, whereas the concept Latin script traditionally covers remains as a category to explain their homological similarities. We can also place it as a macro-script that wraps up its variation, if taken synchronically. This paradigm, on one hand, encourages us to turn our eyes to actual usage and environment within a specific writing system (including interactions between glottic symbols and punctuation, regional variation of handwriting, etc.) rather than imposing the common “Latin script” framework, while on the other hand, draws our attention to the dynamism of historical development and diffusion: from which, and to which, a script tradition of a writing system is transferred, which represents a true richness the ever-evolving Latin-script world.

As for why the idea of a homogenous Latin script has been retained, it is suggested that, paradoxically, it is due to common belief. If one remembers the words of Rogers (2005) cited in Section 3.1, he said: “Western Europe, however, maintained a sense of cultural unity which preserved the Roman alphabet intact.” We would say it is more likely that, the “Roman alphabet” is an artifact of the cultural unity. The shared cultural, religious, and technological background has made people believe in its identity independently of what it is in reality. And it is certainly understandable, because various technologies and social institutions that enable the art of writing play essential roles in actualization and sustainment of each writing system. In this sense, an explanation found in the SIL International website grasps the essence very nicely:

script — a maximal collection of characters used for writing languages or for transcribing linguistic data that share *common characteristics of appearance*, share *a common set of typical behaviours*, have *a common history of development*, and that would be *identified as being related by some community of users*. Examples: Roman (or Latin) script, Arabic script, Cyrillic script, Thai script, Devanagari script, Chinese script, etc. (Lyons et al. 2001; emphasized by the author)

We find that this definition represents a more correct way to capture the current multi-faceted status of this concept. It is not a purely grammatological notion as it may seem, but something influenced by sociological perception, especially at the field site.

This situation is reminiscent of the parlance regarding regional language protection in China. In Europe, the advocates of minority languages are eager to address their systems as “languages,” emphasizing difference with their neighbors, even when they are in the middle of a continuum. The Chinese counterparts, however, keep calling theirs “dialects” even in the most enthusiastic tone. The wording is upheld by a

common cultural belief, which in turn is backed up by their ethnic identity, that the entire spread of Chinese is a single language, although its major “dialects” have little mutual intelligibility.

As we revisit Latin script, what has supported its existence can be likewise named as the greater sociological intervention, or to say, the “common sense,” over the purely grammatological analysis. Regardless of how we are going to cope with Latin script in the future, we strongly believe that we must reappraise the crude reality laid out in front of us concerning its consistency with its value in our theoretical world, rather than simply affirming or repackaging traditional ideas with a new appearance. We also suppose that a similar discussion could be made against other major groups entitled as single “script,” as well as other entities given a name in previous grammatological research. What we discussed in this paper is probably the tip of the iceberg, and much more would be still left hidden.

Lastly, we emphasize the fact that we are not trying to get rid of *script* from the schema defined in Section 1.2, or to incorporate its faculty into another concept. We did not verify whether different writing systems are able to share the same script or not. Topics like whether a concept *script* is valid or useful, and if so, whether it should be subordinated to each writing system or not, are untouched in this paper, though we recognize the importance of such questions that need to be explored in the future. What we have shown at this point is that the alleged vast uniformity of Latin script is unlikely to stand.

5. Conclusion

After having reviewed the current definitions of *script* and its expected nature, our argument is: the entity we call *Latin script* when we use terms *script* and *writing system* to state “(English/French/Indonesian etc.) writing system uses Latin script” is:

- theoretically problematic if regarded as a consistent concept, applied uniformly across writing systems which is supposed to use it
- a socially motivated idea, unlikely to be a valid single script for grammatological analyses
- better positioned as a genealogical grouping or a macro-system (macro-script)

It is expected that similar claims are likewise to be made concerning most cross-regional “scripts,” such as Arabic, Cyrillic, or Chinese. How actually scripts of the world can be alternatively established would be an important task and remains to be seen in the future. We also hope that those categorical entities descended from traditional abstraction should undergo due scrutiny and refinement so that they can be fitted for further academic discussion.

References

- Alessandrini, Jean (1979). “Nouvelle classification typographique: Codex 1980”. In: *Communication et Langages* 43, pp. 35–56.
- Comrie, Bernard (1996). “Languages of Eastern and Southern Europe”. In: *The World’s Writing Systems*. Ed. by Peter T. Daniels and William Bright. Oxford: Oxford University Press.
- Coulmas, Florian (1996). *The Blackwell Encyclopedia of Writing Systems*. Cambridge: Blackwell Publishers.
- (2003). *Writing Systems: An Introduction to Their Linguistic Analysis*. Cambridge University Press.
- Daniels, Peter T. (1996). “The Study of Writing Systems”. In: *The World’s Writing Systems*. Ed. by Peter T. Daniels and William Bright. New York: Oxford University Press.
- (2009). “Grammatology”. In: *The Cambridge Handbook of Literacy*. Ed. by David R. Olson and Nancy Torrance. Cambridge: Cambridge University Press, pp. 25–45.
- (2018). *An Exploration of Writing*. Equinox Publishing.
- Daniels, Peter T. and William Bright, eds. (1996). *The World’s Writing Systems*. Oxford University Press.
- DeFrancis, John (1989). *Visible Speech: The Diverse Oneness of Writing Systems*. University of Hawaii Press.
- Haralambous, Yannis (2007). *Fonts & Encodings*. Sebastopol, CA: O’Reilly.
- Harris, Roy (1995). *Signs of Writing*. New York: Routledge.
- (2000). *Retinking Writing*. New York: Continuum.
- Harvey, Christopher (2009). “SENĆOTEN (Saanich, Northern Straits Salish)”. In: *Languagegeek*. <http://www.languagegeek.com/salishan/sencoten.html>.
- Jamra, Mark (2015). *Phoreus Cherokee*. <https://www.unicode.org/L2/L2015/15214-phoreus-chokeee.pdf>.
- Kimura-mo (2017). “DRAGON 1/72 Me1101”. In: 早くて安くで解像度が低い [Fast, cheap and low resolution]. <http://natsupon.blog60.fc2.com/blog-entry-420.html>.
- Knight, Stan (1996). “Roman Alphabet”. In: *The World’s Writing Systems*. Ed. by Peter T. Daniels and William Bright. Oxford University Press.
- Kobayashi, Akira [小林章] (2012). “c-h 合字・t-z 合字 [c-h ligature and t-z ligature]”. In: 小林章の「タイプディレクターの眼」 [Akira Kobayashi’s “Type Director’s Eye”]. <http://blog.excite.co.jp/t-director/17658486/>.
- Kohrt, Manfred (1985). *Problemgeschichte des Graphembegriffs und des frühen Phonembegriffs*. Tübingen: Niemeyer.
- Lockwood, David G. (2001). “Phoneme and Grapheme: How Parallel Can They Be”. In: *LACUS Forum* 27, pp. 307–316. http://www.lacus.org/volumes/27/404_lockwood_d.pdf.
- Lyons, Melinda et al. (2001). “Glossary”. In: *NRSI: Computers & Writing Systems*. <http://scripts.sil.org/Glossary>.

- My another account (2014). *Car of Polish City Guard (Straż Miejska), in Warsaw Old Town*. https://commons.wikimedia.org/wiki/File:Stras%C5%BC_Miejska.JPG.
- Özer, Nermin Özcan (2016). “Öğretmen Adaylarının Bitişik Eğik El Yazısı Alfabeti İle İlgili Düzeyleri [Levels of Student Candidates regarding Script Italic Handwriting Alphabet]”. In: *Western Anatolia Journal of Educational Sciences INOVED 2016*, pp. 199–224. http://webb.deu.edu.tr/baed/giris/baed/inoved_12.pdf.
- Pike, Kenneth L. (1954). *Language in Relation to a Unified Theory of the Structure of Human Behavior*. Glendale, CA: SIL International. <http://hdl.handle.net/2027/mdp.39015004829167>.
- Rogers, Henry (2005). *Writing Systems: A Linguistic Approach*. Hoboken, NJ: Wiley.
- Sampson, Geoffrey (1985). *Writing Systems: A Linguistic Introduction*. Stanford, CA: Stanford University Press.
- (2015). *Writing Systems*. 2nd ed. Sheffield: Equinox Publishing Ltd.
- SIL International (n.d.). “Latin”. ScriptSource, <http://scriptsource.org/scr/Latn>.
- Sproat, Richard (2000). *A Computational Theory of Writing Systems*. Cambridge: Cambridge University Press.
- Taylor, John R. (1995). *Linguistic Categorization: Prototypes in Linguistic Theory*. 2nd ed. Oxford: Clarendon Press, Oxford University Press.
- The Unicode Consortium (2016). *The Unicode Standard, Version 9.0.0*. The Unicode Consortium.
- (n.d.). “Glossary of Unicode Terms”. <https://unicode.org/glossary/>.
- Wittgenstein, Ludwig (2009). *Philosophische Untersuchungen = Philosophical Investigations*. Trans. by G. E. M. Anscombe, P. M. S. Hacker, and Joachim Schulte. 4th ed. Oxford: Wiley-Blackwell.
- XYZ4096 (2015). “pic.twitter.com/xa3GdqUdSu”. <https://twitter.com/XYZ4096/status/651742512881664001>.

Digraphia: the Story of a Sociolinguistic Term

Sveva Elti di Rodeano

Abstract. Digraphia is a metalinguistic term referring to the coexistence of two scripts for one language. The main objective of this paper is to provide the all the attestations of the term “digraphia” and its meanings, in order to propose an unambiguous definition for Grapholinguistic Studies.

This reflection has a twofold aim: it is meant to establish a definition for this term, useful for a linguistics and language glossary; and it is intended to highlight the necessity of a precise designation in the Grapholinguistic field. The terminology is in fact the *trait d'union* between the disciplines involved (computer science, statistics, linguistics, sociology, etc.). The interdisciplinarity of terminology results from the character of the designations: they are linguistic items, conceptual elements and vehicles of communication. In order to spread, share and improve knowledge in this special field,¹ the communication between scholars must be normalized and the concepts must be standardized.

1. Introduction

The metalinguistic reflection has spread the generally accepted opinion that “language and writing are two distinct systems of signs; the second exists for the sole purpose of representing the first” (Saussure, 1959,

It is a pleasant duty to express my gratitude to my supervisor Prof.ssa Raffaella Bombi (Udine), who spent time with me discussing this paper, and Prof. Vincenzo Orioles (Udine), who inspired me the idea of dealing with metalinguistic terminology of writing systems. All remaining errors are my own responsibility.

Sveva Eltidi Rodeano
University of Udine
Dipartimento di Studi Umanistici e del Patrimonio Culturale
vicolo Florio, 2/b
33100 Udine, Italy

1. It is relevant to note that even the name of this field is still uncertain. English scholars use “graphematic, graphematik, grapholinguistics,” German “Schriftlinguistik, Grapholinguistik, Graphemik,” Italian “grafemica, grafematica”.

The designation “Schriftlinguistik” is relatively new, Dürscheid (2006, p. 12) refers to Nerius and Augst (1988) as its first user.

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1.

Fluxus Editions, Brest, 2019, p. 111–126. <https://doi.org/10.36824/2018-graf-elti>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

pp. 23–24). Therefore, there is not a linguistic conception of writing system, since the *écriture* is just the portrayal of the *langue*.

The *fil rouge* and common thread that wends its way through this paper is whether writing systems are strictly symbolic systems representing languages, or linguistic entities themselves.

For this discussion, which goes beyond this paper but it seems impossible not to mention it at all, it is useful remind us of Aristotle's definition of writing (Περὶ Ἑρμηνείας 16a 3–9), since it became axiomatic in the Western tradition.

Ἔστι μὲν οὖν τὰ ἐν τῇ φωνῇ τῶν ἐν τῇ ψυχῇ παθημάτων σύμβολα, καὶ τὰ γραφόμενα τῶν ἐν τῇ φωνῇ. καὶ ὥσπερ οὐδὲ γράμματα πᾶσι τὰ αὐτά, οὐδὲ φωναὶ αἱ αὐταί· ὧν μέντοι ταῦτα σημεῖα πρώτων, ταῦτα πᾶσι παθήματα τῆς ψυχῆς, καὶ ὧν ταῦτα ὁμοιώματα πράγματα ἦδη ταῦτά.

Now spoken sounds are symbols of affections in the soul, and written marks symbols of spoken sounds. And just as written marks are not the same for all men, neither are spoken sounds. But what these are in the first place signs of—affections of the soul—are the same for all; and what these affections are likenesses of—actual things—are also the same. (Acrill, 1991, p. 25)

Here τὰ γραφόμενα, written words, are symbols, σύμβολα, of spoken words, while spoken words are symbols of affection of the soul. Therefore, the relationship between written words, spoken words, words and things are linear and monodirectional.

Currently, even if several books have been published that suggested the equity of writing with speech sound (Massias 1828, Assmann 1991), writing became a device for expressing language, rather than being the mere representation of speech from its origin (Gelb, 1963, p. 13). This idea leaves space for recognizing non linguistic functions of writing, but, since the aim of this paper is focused on the metalinguistic aspect of writing, I will intentionally go beyond this and consider writing within the framework of its respective languages.

Indeed, sociolinguistic studies have seeded the soil where we can now find different approaches to the study of writing systems: the idea that scripts are able to modify a community of speakers is now disseminated, since they are bound to religious motions, identity claims, and even scientific progress. One of the most eminent scholars in this field, Florian Coulmas, stated that:

rather than being mere instruments of a practical nature, they [scripts and orthographies] are symbolic systems of great social significance which may, moreover, have profound effect on the social structure of a speech community” (Coulmas, 1989, p. 226).

Even though nowadays several scholars have seen a utilitarian relation between language and writing, because writing is “not language [...] writing does represent language” (Rogers, 2005, p. 2), where language

must be interpreted as phonetic representation, various studies have been published in recent years regarding adoption/change of script(s), no linguistic dictionary has yet recorded the lexicon referring to these phenomena. For instance, Bußmann and Cotticelli Kurras (2007, p. 202) offered the following explanation for the term “digraphia”:

digrafia

[gr. gráphein ‘scrivere’].

Rappresentazione di un fonema tramite due segni grafici, ad es. ingl. <sh> per [ʃ], ted. <ch> per [x] o [ç]. I due segni interessati costituiscono nella loro unità un “digrafo”.²

In this paper, I illustrate the history of one most attested words in the literature of writing systems, as “digraphia,”³ considering it as a typology, as well as diglossia (Ferguson, 1963, p. 163), useful both for graphemic studies, since it strictly refers to scripts, and both for linguistic studies, since it considers the coexistence of two scripts for one language. The term is intended as a sociolinguistic typology, in comparison with its linguistic parallel “diglossia,” therefore its attestations and history of meaning are divided into ante- and post- adventum of Charles Ferguson’s contribution.

Above all, “digraphia” is a metalinguistic term and for this field afterwards I will propose an unambiguous definition.

2. Digraphia Ante Diglossia: The Pioneers

The adjective “digraphic” appeared for the first time in the narration written by Demetrios Pierides, a bank manager in Larnaca and collector of classical antiquities. Pierides recounted his discovery of an inscription with the same text written down in two different scripts, Greek and Cypriot:

In the summer of 1873 I became possessed of an inscription in Greek and Cypriote, then discovered in Larnaca, the ancient Citium. [...] As the language is the same in both parts, and only the writing differs, I prefer calling this inscription *digraphic*, instead of *bilingual*. (Pierides, 1875, p. 38)

2. “digrafia//[gr. gráphein ‘to write’].//The representation of a single phoneme with two graphic signs, e.g. Engl. <sh> for [ʃ], Germ. <ch> for [x] or [ç]. The two written signs constitute a “digraph”.”

Similar explanations could be found in Pei and Gaynor (1954, p. 57), Hartmann and Stork (1972, p. 67), Mackay (1989, p. 159), Trask (1996, p. 113): 113), Beccaria (1994, p. 230), Matthews (1997, p. 98), Bußmann (1998, p. 315), Crystal (2008, p. 145).

3. The history of this word is briefly written down in Britto (1986, pp. 309–310), Grivelet (2001, pp. 1–6) and more extensively in Bunčić, Lippert, and Rabus (2016, pp. 27–50).

It is highly unlikely that Pierides had been trained in linguistics, but surely he was an enthusiast and amateur of the Greek world, therefore it is likely that he had formed the word “digraphic” with the Greek prefix δι- “two, double,” lacking the functional distribution sense.

In the same year, the numismatist Alfred von Sallet used *zweischriftig* to describe the inscription he recovered from some Cypriot coins:

[...] einige dieser Münzen, welche als zweischriftig—sit venia verbo—besonders interessant sind, geben neben der cyprischen auch die griechische Legende [...].⁴ (Sallet, 1875, p. 132)

He found that two different scripts (Cypriot and Greek) were employed, an instance for which he begged our pardon but no words appeared more suitable to describe it than “zweischriftig,” which is a new formation in the absence of an appropriate metalinguistic word.

In the same way as Pierides, the orientalist Joseph Halévy used *digraphique* in the description of some Sumero-Akkadian inscriptions, disagreeing with the assiriologists who called them “bilingual”.

Les textes réputés bilingues de l’antique Babylonie, quel que soit leur caractère, ne peuvent donc être que des rédactions digraphiques exprimant une langue unique, l’assyrien.⁵ (Halévy, 1883, p. 255)

Evidently, Halévy’s critique was justified by his conviction that Sumerian was not a language at all, but just an alternative script for Akkadian. This was the reason for it being not worth mentioning his contribution to the diffusion of the term “digraphia,” but just to be through regarding its attestations.

3. Digraphia Post Diglossia: The First Attempts

In 1959, Charles Ferguson introduced the concept of Diglossia as

a relatively stable language situation in which, in addition to the primary dialects of the language (which may include a standard or regional standards), there is a very divergent, highly codified (often grammatically more complex) superposed variety, the vehicle of a large and respected body of written literature, either of an earlier period or in another speech community, which is learned largely by formal education and is used for most written and formal spoken purposes but is not used by any sector of the community for ordinary conversation. (Ferguson, 1959, p. 336)

4. “[...] some of these coins, which are especially interesting as they are zweischriftig—if you pardon the expression—, also give the Greek legend next to the Cypriot [...]”.

5. “The texts of ancient Babylonia regarded as bilingual, whatever their nature, can therefore only be digraphic recensions conveying a single language, Assyrian.”

Thus, the digraphic sense of functional distribution refers to this definition. Immediately following Ferguson, digraphia has been perceived as a sociolinguistic typology.

The linguist Robert Lafont, describing the features of Occitan, which reveal both two linguistic varieties (French and Occitan) and two graphic varieties (classic and mistral), wrote:

La situation se présente donc ainsi: deux langues d'expression écrite, mais l'une est populaire, familière, et en tout état de cause, dominée par l'autre. On ne parvient à l'écrit occitan que si l'on a déjà appris à lire en français. [...] La situation de diglossie occitane n'est donc pas semblable absolument à celles qu'on peut trouver en d'autres lieux de contacts linguistiques: elle se complète par une situation de digraphie.⁶ (Lafont, 1971, p. 95)

Though the reference to Ferguson was more than evident, it is not clear if Lafont conceived digraphia as a functional distribution of scripts, because here we are also dealing with orthographies.⁷

Otherwise, Petr Zima, describing the functional distribution of the two scripts (Latin and Arabic alphabets) in which Hausa is attested, clearly referred to Ferguson's diglossia. Furthermore, he discerned whether we can talk about a case of "digraphia" or a case of "diorthographia," phenomenon just now observed in Lafont:

Digraphia: "Two types of written form of one language co-exist, based upon the usage of two distinct graphical systems (scripts) by the respective language community."

Diorthographia: "Two types of written form of a particular language co-exist, using the same script, but they are based upon the usage of two distinct orthographies by the same language community." (Zima, 1974, p. 58)

Regarding this distinction, Paul Wexler had already received Ferguson's lesson, but he still did not have an appropriate terminology: therefore, he spoke about "orthographic diglossia" as a case of diglossia in which "different scripts may be used by a single ethnic group for different purposes [...]" (Wexler, 1971, p. 340).

For these reasons, we can see how the concept of a functional distribution between two linguistic varieties, in this case writing varieties,⁸

6. "The situation is as follows: two written languages, one of which is vernacular, familiar, and at any rate, dominated by the other. [...] Therefore the Occitan situation of diglossia does not resemble those which one can find in other places of language contacts at all: it is completed by a situation of digraphia."

7. Likewise, the authorship of "diglossia" (generally ascribed to Charles Ferguson, even if the term was used before him in 1885 by Emmanuel Roidis and Jean Psichari), the one of "digraphia" has been ascribed to Zima by Grivelet (2001, p. 1) and to Lafont by Bunčić, Lippert, and Rabus (2016, p. 40).

8. Having said that, I consider writing systems as a linguistic category.

has already been perceived by scholars, but there was not a specific and unambiguous terminology.

Just two years after Zima, the anthropologist James R. Jaquith wrote two papers regarding discordant orthographies in advertisements, and he defined digraphia as “the graphic analog of diglossia” (1996, p. 303). Another clue that the term had not still been accepted was that the title of the paper itself was wrongly spelled as “diagraphia”.⁹

3.1. Digraphia Post Diglossia: The Role of Giorgio Raimondo Cardona

Since 1978, G. R. Cardona had wondered “quanto dei concetti propri della sociolinguistica e della etnografia della comunicazione può essere estesi alla scrittura?” (1978, p. 67). The same question appears in *ILuoghi del sapere* (185), but it seems that for him the matter was already solved, given the paragraph “sociologia della scrittura” (1981), where he evidently used sociolinguistic categories—for instance writing-community (speech-community), prestige, and writing repertoire—for writing systems:

I vari insieme coerenti di simboli posseduti dallo scrivente possono essere paragonati, con le precisazioni che si diranno, alle varietà linguistiche che formano il repertorio verbale. Analogamente a quando si dà nell'uso di questo repertorio, anche nello scrivere si sceglierà la varietà scrittoria più adatta all'evento scrittorio.¹⁰ (ibid., p. 103)

He also added:

In ogni società le varie produzioni simboliche si diversificano e si strutturano in modo funzionale alla società stessa [...] le differenze verranno sempre rese funzionali agli scopi della società. [...] Dove, tra le varie forme di produzione simbolica, compaia la scrittura, questa non potrà certo costituire eccezione, ma sarà soggetta all'esigenza modellizzante propria della cultura.¹¹ (ibid., p. 89)

Having said that, it is now possible to reinterpret the following paragraph, where Cardona illustrated the spread of the Arabic script:

9. The same fate occurred to Dale (1980) and Collin (2005).

10. “The coherent sets of signs owned by the writer can be compared, with the previous mentioned below, with the linguistic varieties of spoken language. Likewise in the case of spoken language, in the written language the variety must be chosen in accordance with the written occasion.”

11. “In every society, different symbolic representations are functionally diversified and structured, in accordance to each society [...] the differences will be made functional to the purposes of the society. [...] If writing appears among the various forms of symbolic representation, then this certainly cannot constitute an exception, but will be subject to the modeling requirement of culture itself.”

La scrittura araba si è diffusa, nel giro di qualche secolo, [...] in Spagna, in Sicilia, in Serbia però essa, se è stata certo in uso spesso per qualche secolo per documenti e scritture religiose, non ha attecchito nemmeno nel periodo in cui gli Arabi erano effettivamente padroni della situazione. Possiamo pensare che in Sicilia vi sia stato un lungo periodo addirittura di bilinguismo siciliano arabo, ma non si può dire altrettanto di un periodo di digrafia latino-araba.¹²
(*ibid.*, pp. 128–129)

Even if he did not explicitly write which ones could be the functions of, respectively, the Arabic script and the Latin script, due to what we have just examined, Cardona meant “digrafia” as coexistence of two writing varieties sorted in different purposes.

Shortly afterwards, another Italian scholar, Carlo Consani, published a trilogy of papers titled “Bilinguismo, diglossia e digrafia nella Grecia antica” (1988–1990), in which, drawing on the philological current led by Pierides, he used the concept of digrafia as a sociolinguistic typology useful for describing the case of Cyprus. There, diglossia and digraphia coexisted, since the scripts implied are two: the Greek alphabet and Cypriot syllabary:

[...] tutti questi elementi mostrano a quali drastiche restrizioni, ai diversi livelli diatopico, diastratico e diafasico-situazionale, risponda l’uso del dialetto e della scrittura sillabica.¹³
(Consani, 1990, p. 77)

Therefore, the meaning of digraphia here is perfectly in step with Ferguson’s shared influence post-1959.

3.2. Digraphia Post Diglossia: New Models

In the last decade of the twentieth century, the functional distribution is an essential part of the digraphia concept. As a matter of fact, based on the case study of Serbo-Croatian and Japanese, several new models emerged. In Haarmann (1993, pp. 153–154), digraphia is employed in the description of a Korean text, where the use of two different scripts shows a functional distribution depended on their prestige. At the same time, he used bigraphism talking about languages, like Serbo-Croatian and Japanese, which do not show any differences in terms of prestige between their writing systems (2006, pp. 2406–2407).

12. “The Arabic writing has spread, during some centuries, [...] in Spain, Sicily and Serbia, it has been used in religious documents, but it did not hold even when Arabs were in charge. We can guess that there has been a long Sicilian-Arabic bilingualism period, but it can not be said about a digrafia Latin-Arabic period.”

13. “[...] all these elements are illustrative of the diastratic conditions, from the diastratic, diatopic and diafasic point of view, under which dialect and syllabic writing are used.”

In parallel to Ferguson's model, Chiang proposed a similar one more suited to digraphia, since it distinguished more analytically how the distribution of scripts does work: he related more prestigious scripts to "semantic scripts," while less prestigious scripts are related to "phonetic scripts".

It [the H variety] is retained to represent the same meaning although it no longer represents the pronunciation accurately. I call the script thus used the semantic variety. For instance, in English, "night" used to be pronounced /nixt/. This script form was preserved after the pronunciation had changed to /nait/. (Chiang, 1995, p. 112)

My use of the terms "semantic" and "sound" here are similar in meaning to Haas' use of the terms "pleremic" and "cenemic". See William Haas, "Determine the Level of a Script". (*ivi*, 125).

From the point of view of a speaker whose writing system does not accurately represent the phonetic form of language, this distinction turns out to be interesting. Orthographia, as "the correct graphia," is not necessarily phonetic, it is rather more common for the contrary: in less controlled contexts, where a less prestigious script is used, there could be orthographic mistakes; meanwhile in more controlled contexts, where a more prestigious script is used, there should not be orthographic mistakes.

3.3. (Implicit) Digraphia Post Diglossia

Since a scientific terminology for the contact between scripts has not been established, many scholars explain examples of digraphia, without mentioning the term. Among them, Sergio Pernigotti, talking about Ancient Egypt's writing systems (hieroglyphs and hieratic), wrote:

Questa situazione della contemporanea presenza di due scritture strettamente correlate tra di loro dal punto di vista della grafia e distinte soltanto nel loro uso di mantenne sostanzialmente immutata dall'unificazione del paese fino a circa l'inizio del VII secolo a.C., quando questo panorama abbastanza semplice venne complicato dall'introduzione di un terzo tipo di scrittura che si affiancò ai due precedenti e che noi chiamiamo ancora oggi con il termine, coniato da Erodoto, di "demotica" [...] ¹⁴ (Pernigotti, 1986, p. 30)

14. "The simultaneous presence of two scripts, which were interrelated from the graphic point of view and separated by their use, remained stable from the unification of the country until about VII BC., when it was introduced the third script, which was accompanied by the previous two and which now is called "demotic," as Herodotus coined first."

This case could be found in Bunčić, Lippert, and Rabus (2016, pp. 256–275), as an example of “bigraphism,” while, the later co-presence of three scripts (hieroglyphs, hieratic, and demotic) is called “scriptal pluricentricity” (*ivi*, 183–185).

In my opinion, this case could be a perfect example of the so-called “diachronic digraphia” (Berlanda, 2006, pp. 89–98): the internal differences in style of Egyptian Hieroglyphs led up to an internal digraphia,¹⁵ developing two different scripts which are used in two different domains, hieroglyphs for monumental inscriptions, hieratic for administrative documents.

Berlanda (*ivi*, 72–73) wondered why these two other scripts have evolved and suggested that the reason lay in the necessity of keeping it [the hieroglyphs] “clean from change”.

This idea, implying that a new script evolved from another one in order to keep the previous one unchanged (in terms of shape, domain of use, users) has a twofold implication: (A) uncommon domains of use allow the user to change the style of the script; (B) the user considers the script in charge suitable for the uncommon domain.

The consequence of (A) is that, considering Ancient Egypt’s scripts, if the calligraphy of hieroglyphs had been modified until becoming hieratic, we should have found some attestation of an intermediate phase between hieroglyphs and hieratic graphemes, and we have not. The consequence of (B) is that we could have found some attestation of hieroglyphs used in non-monumental texts, and we have not. For these reasons, it is more likely that the reason for the creation of one and then another script lies in the spread of use of writing, from the point of view of users (in Egypt, at first, it was prerogative of scribes) and of domains (at first used only in monumental texts).

Again, about hieroglyphs, Louis Godart described the scripts attested in II millennium AC in Crete:

A priori, sarebbe logico supporre che i motivi dell’esistenza di due scritture diverse nella Creta protopalaziale fossero legati alla presenza di popolazioni diverse, che parlavano due lingue diverse e quindi utilizzavano due sistemi grafici diversi [...] La semplicità e la logica di questa ipotesi non sembrano tuttavia resistere a un esame più attento dei dati archeologici. [...] Se si esamina la cronologia della scrittura geroglifica e si analizzano i tipi di supporti sui quali le prima due scritture cretesi sono attestate, si notano alcuni elementi sorprendenti:

1. I primi documenti geroglifici rinvenuti sono i sigilli, che recano testimonianza della cosiddetta scrittura di Arkhanes, mentre i più antichi documenti d’archivio in nostro possesso sono le tavolette in lineare A di Festo,

15. Hieratic script is the cursive version of hieroglyphic script, and the later demotic script is even more cursive than the hieratic one. Dale (1980, p. 6): 6) called this case “internal di[a]graphia”.

i documenti del Deposito geroglifico di Cnosso e gli archivi in geroglifico del Quartier Mu di Mallia.

2. La maggior parte dei documenti scritti in geroglifico è costituita da sigilli o impronte di sigilli, mentre non un solo documento in lineare A ci è pervenuto su questo tipo di supporto.

Una tale convergenza di dati non può essere frutto del caso, e forse possiamo trovare in questa differenziazione del supporto delle scritte i motivi che hanno spinto gli amministratori palaziali minoici dell'inizio del II millennio a.C. a utilizzare i due sistemi di scrittura di cui stiamo trattando.¹⁶

(Godart, 1992, pp. 139–140)

This is an example for which Bunčić, Lippert, and Rabus (2016, p. 58) would have used the definition “medial digraphia,” whereas the only discriminating factor for the choice of one script rather than the other one is the medium scriptiois. In this case, we are dealing with a case of coexistence of two scripts and, fortunately, their distinct material vehicle.

As detailed below, Massimiliano Marazzi described more extensively the reason for the coexistence of two scripts in the case in the Anatolia of II millennium A.C., hence the coexistence of cuneiform and Anatolian hieroglyphs:

[...] si assiste allo sviluppo contemporaneo e parallelo di 2 sistemi scrittori, facenti certamente capo agli stessi ambienti scribali e inizialmente profondamente diversi e differenziantisi quanto a:

- scelta dei supporti;
- principi di organizzazione dei segni su supporti stessi;
- tracciato, articolazione e organizzazione dei grafemi che compongono il sistema;
- rapporto fra codice scrittorio e codice linguistico.

[...] Se infatti il sistema cuneiforme hittita si afferma quale sistema scrittorio lineare fonetico su tavoletta d'argilla, utilizzato per esprimere in tutti gli ambiti [...] le lingue correnti dell'epoca [...], il cd. “geroglifico anatolico” rimane limitato alla sola superficie glittica, mantenendo intatto l'impianto originario di sistema segnico, organizzatosi essenzialmente all'interno di uno spazio, quello appunto del sigillo [...]¹⁷

(Marazzi, 2009, pp. 116–117)

16. “A priori, it would be logical assume that the reasons two different scripts exist in Protopalatial Crete were related to the presence of different peoples, speaking different languages and writing in different scripts [...] This simple and logic approach can not withstand an accurate archaeological study. [...] A careful analysis of both the chronology of the hieroglyphic script and the supports of the first two Cretan scripts displays some surprising elements://1. The first hieroglyphic documents found are seals, which bear witness of the so called Arkhanes script, while the more ancient archival documents are the linear A tablets from Phaistos, Knossos and Mallia.//2. The majority of the hieroglyphic documents are seals and seal imprints, while not a single linear A documents appear on this kind of support. The convergence of information can not be a coincidence, and perhaps we can find the reasons why two scripts have been used in the 2nd millennium BC in the distinction of support.”

17. “There is the development of two scripts, related to the same scribal environment and deeply different for://– support selection//– principles of signs

In this case we have the coexistence of two different scripts, which differ in the way of organizing their signs. Hittite cuneiforms consist of a large part of syllabic signs, whereas Anatolian hieroglyphs could have phonetic value or be direct symbol of a thing. Saying that hieroglyphs could be direct symbol implies that they could not rely on any language. From these premises, we should add that the “scribal habits” were highly sophisticated and multilinguistic. At last, in my opinion, this example offers us the occasion to highlight the role of the medium script: it could not be a coincidence that the first supports on which we have found Anatolian hieroglyphs are seals, thing characterizing by peculiar use and shape, and which perfectly fits for visual composition, rather than textual ones.

4. Digraphia Nowadays

Recently two volumes have been published regarding contact between writing systems: *Biscriptality, a Sociolinguistic typology*, by Bunčić, Lippert, and Rabus (2016) and *Contatti di lingue, contatti di scritture* by Baglioni and Tribulato (2015). The first one, chronologically successive, illustrates a theory which Italian scholars have been already been aware of: an heuristic model which investigates all the so called “biscriptality” cases (*digraphia, diglyphia, diorthographia, bigraphism, biglyphism, biorthographism, scriptal pluricentricity, glyphic pluricentricity, orthographic pluricentricity*), identified on the basis of two axes, the sociolinguistic one, which describes the relation between the two scripts (privative, equipollent, and diasituative) and the graphematic one, which identifies three levels of distinctions (script, glyphic variation, orthography). The only one called “digraphia” is when there is a clear functional division of domains of the considered script:

Digraphia with scripts in a privative opposition, based on a diaphasic, diastatic, diamesic or medial distribution.

(Bunčić, Lippert, and Rabus, 2016, p. 62)

On the other side, in Baglioni-Tribulato, even if they explicitly referred to the previous text, the concept of digraphia appears from a different point of view, given that “di questo macro fenomeno, poi, è possibile riconoscere diverse manifestazioni sociali, a seconda che le due

organization//– layout, structure and organization of signs//– relation between writing and language//[...] While the Hittite Cuneiform asserted its linear and phonetic scripts on clay tablets, and it was used in all context of use [...] the languages of that time [...]. the so called “Anatolian hieroglyphs” remains restricted only to the engraved surface, maintaining the original signs organization intact, which was mainly created within the seal written space [...].”

scritture abbiano un prestigio diverso e conoscano pertanto una netta ripartizione funzionale” (Baglioni and Tribulato, 2015, p. 15). It seems that the model of Bunčić-Lippert-Rabus has been interpreted too *in strictu sensu*: the only one remaining discriminating factor for the choice between one or the other scripts is the prestige of them. Again Baglioni-Tribulato specify:

differenza della diglossia, la ‘digrafia’ è una condizione non dell’intero repertorio, ma di una lingua specifica considerata nella sua relazione con la scrittura: ne consegue che diglossiche sono le società, mentre ‘digrafiche’ sono le lingue; nella diglossia le due lingue sono l’una gerarchicamente subordinata all’altra, mentre nella ‘digrafia’ [...] non sempre è individuabile una ripartizione funzionale delle due scritture [...]; esistono lingue per la cui notazione sono o sono stati impiegati più di due sistemi di scrittura e per le quali pertanto l’etichetta di ‘digrafia’ non è utilizzabile”¹⁸ (ibid., p. 14)

Therefore, here we have another point of view of the total subject: as well as speakers’ communities being diglossic, languages are digraphic, which do not always show the criterion with their preference for one script. For this reason, it seems pointless to compare this volume to the previous one, since the starting point of the matter is completely different: Baglioni-Tribulato consider writing system a feature of language, while Bunčić-Lippert-Rabus consider it as a linguistic object worthy of research per se.

In fact, they clarify:

By analogy with diglossia, in some cases of digraphia it even makes sense to speak of an H writing system and an L writing system. [...] However, there are also many cases of digraphia in which the feature governing the privative opposition does not lend itself to a high-low distinction. [...] The use of two writing systems for one language is a case of linguistic variation. Therefore, it seems appropriate to use the well-known model of linguistic variation assembled by Coseriu (1992, pp. 280–292), consisting of diachronic, diatopic, diastatic and diaphasic variation. (Bunčić, Lippert, and Rabus, 2016, p. 57)

Here, as Coulmas said, variation of writing systems is a linguistic variation, which means that writing systems must be studied as linguistic objects.

In summary, nowadays there are two understandings of digraphia: one, specular to the notion of diglossia, interprets it as a linguistic

18. “Unlike diglossia, digraphia is a condition of the specific language, given in its relation with the writing: it follows that societies are diglossic, while languages are digraphic; in diglossia the languages are hierarchically subordinated, while in digraphia [...] the functional distribution is not always identifiable [...]; there are languages which are, or have been, written with more than two scripts, for which “digraphia” is not suitable”.

change, the use of two scripts in privative opposition for the same language, where variations can be identified variations, like diaphasic, diastratic, or diamesic ones; the other provides the notion of prestige as the single determining factor for the choice between two scripts.

The resulting definition of “digraphia” should include all the previous illustrated ideas, but in this way it could not fit a dictionary in terms of unambiguity and coherence, even if the price is the general vagueness. In my opinion it will not be wrong to add the following to Bußmann-Cotticelli Kurras’ definition:

Digraphia: a sociolinguistic typology used for describing writing system contact in a speech community where several factors (diaphasic, diastratic, and diamesic) lead in the choice between scripts.

References

- Acrill, John (1991). “Translation of Aristotle’s *De Interpretatione*.” In: *The Complete Works of Aristotle*. Ed. by Jonathan Barnes. Princeton: Princeton University Press.
- Assmann, Jan (1991). *Stein und Zeit: Mensch und Gesellschaft im alten Ägypten*. Munich: Wilhelm Fink.
- Baglioni, Daniele and Olga Tribulato (2015). *Contatti di lingue – contatti di scritture. Multilinguismo e multigrafismo dal Vicino Oriente Antico alla Cina contemporanea [Language Contacts—Writing Contacts. Multilinguism and Multigraphism from the Ancient Near Middle East to Nowadays China]*. Venice: Ca’ Foscari.
- Beccaria, Gian Luigi (1994). *Dizionario di linguistica [Dictionary of Linguistics]*. Turin: Giulio Einaudi.
- Berlanda, Elena (2006). “New Perspectives on Digraphia: A Framework for the Sociolinguistics of Writing Systems”. https://www.omniglot.com/language/articles/digraphia/digraphia_EBerlanda.pdf.
- Britto, Francis (1986). *Diglossia: A Study of the Theory with Application to Tamil*. Washington, DC: Georgetown University Press.
- Bunčić, Daniel, Sarah L. Lippert, and Achim Rabus (2016). *Biscriptality, a Sociolinguistic Typology*. Heidelberg: Universitätsverlag Winter.
- Bußmann, Hadumod (1998). *Routledge Dictionary of Language and Linguistics*. London, New York: Routledge.
- Bußmann, Hadumod and Paola Cotticelli Kurras (2007). *Lessico di Linguistica [Dictionary of Linguistics]*. Alessandria: dell’Orso.
- Cardona, Giorgio Raimondo (1978). “Per una teoria integrata della scrittura [For an integrated study of writing]”. In: *Alfabetismo e cultura scritta nella storia della società italiana, Atti del Seminario tenutosi a Perugia il 29–30 marzo 1977 [Alphabetism and Written Culture in the History of Italian Society. Proceedings of a Seminar Held in Perugia on March 29–30, 1977]*, pp. 51–76.

- Cardona, Giorgio Raimondo (1981). *Antropologia della scrittura* [*Anthropology of writing*]. Turin: Loescher.
- (1990). *I linguaggi del sapere* [*Languages of knowledge*]. Rome: Laterza.
- Chiang, William Wei (1995). *"We Two Know the Script; We Have Become Good Friends": Linguistic and Social Aspects of the Women's Script Literacy in Southern Hunan, China*. Lanham, New York, London: University Press of America.
- Collin, Richard Olivier (2005). "Revolutionary Scripts: The Politics of Writing Systems". In: *iOmniglot: Writing Systems and Languages of the World*. Ed. by Simon Ager.
- Consani, Carlo (1988). "Bilinguismo, diglossia e digrafia nella Grecia antica I: Considerazioni sulle iscrizioni bilingui di Cipro [Bilingualism, Diglossia and Digraphia in Ancient Greece I: Writings on Bilingual Inscriptions from Cyprus]". In: *Bilinguismo e biculturalismo nel mondo antico: Atti del Colloquio interdisciplinare tenuto a Pisa il 28 e 29 settembre 1987* [*Bilingualism and Biculturalism in the Ancient World: Proceedings of an Interdisciplinary Colloquium Held in Pisa on September 28-29, 1987*]. Ed. by Enrico Campanile, Giorgio R. Cardona, and Romano Lazzeroni. Pisa: Giardini.
- (1989). *Bilinguismo, diglossia e digrafia nella Grecia antica II: Le lettere di Filippo V e i decreti di Larissa* [*Bilingualism, Diglossia and Digraphia in Ancient Greece II: The Letters of Philip v and the Larissa Decrees*] (Schwyzer, DGEEP, 590). Vol. 11, pp. 137-159.
- (1990). *Bilinguismo, diglossia e digrafia nella Grecia antica III: Le iscrizioni digrafe cipriote* [*Bilingualism, Diglossia and Digraphia in Ancient Greece III: Cypriot Inscriptions in Two Writing Systems*]. Vol. 25, pp. 63-79.
- Coseriu, Eugenio (1992). "Prinzipien der Sprachgeschichte: Vorlesung im Wintersemester 1990/91 an der Eberhard-Karls-Universität Tübingen".
- Coulmas, F. (1989). *The Writing Systems of the World*. Oxford: Oxford University Press.
- Crystal, David (2008). *A Dictionary of Linguistics and Phonetics*. London: Wiley-Blackwell.
- Dale, Ian R. H. (1980). "Di[a]graphia". In: *International Journal of the Sociology of Language* 26, pp. 5-13.
- Dürscheid, Christa (2006). *Einführung in die Schriftlinguistik*. Göttingen: UTB.
- Ferguson, Charles (1959). "Diglossia". In: *Word* 15, pp. 325-340.
- (1963). "Introduction [to National Languages and Diglossia]". In: *Report of the Thirteenth Annual Round Table Meeting on Linguistics and Language Studies*. Ed. by E. Woodworthand and R. Di Pietro. Vol. 15. Washington, DC: Georgetown University Press.
- Gelb, Ignace Jay (1963). *A Study of Writing*. Chicago: University of Chicago Press.

- Godart, Louis (1992). *L'invenzione della scrittura: dal Nilo alla Grecia* [*The invention of writing: from the Nile to Greece*]. Turin: Giulio Einaudi.
- Grivelet, Stéphane (2001). "Introduction". In: *International Journal of the Sociology of Language* 150, pp. 1–10.
- Haarmann, Harald (1993). "The Emergence of the Korean Script as a Symbol of Korean Identity". In: *The Earliest Stage of Language Planning: The 'First Congress' Phenomenon*. Ed. by Joshua A. Fishman. Berlin: de Gruyter, pp. 143–157.
- (2006). "Language Planning: Graphization and the Development of Writing Systems". In: *Sociolinguistics: An International Handbook of the Science of Language and Society*. Ed. by Ammon Ulrich et al. Vol. 3. Berlin: de Gruyter, pp. 2402–2420.
- Halévy, Joseph (1883). "Étude sur les documents philologiques assyriens". In: *Mélanges de critique et d'histoire relatifs aux peuples sémitiques*. Paris: Maisonneuve et C^{ie}, pp. 241–364.
- Hartmann, Reinhard R.K. and F.C. Stork (1972). *Dictionary of Language and Linguistics*. London: Halsted Press.
- Lafont, Robert (1971). "Un problème de culpabilité sociologique: La diglossie franco-occitane". In: *Langue française* 9, pp. 93–99.
- Mackay, Ian (1989). *Phonetics and Speech Science. A Bilingual Dictionary*. New York, Bern: Peter Lang.
- Marazzi, Massimiliano (2009). *Lineare o geroglifico? Sistemi scrittori a confronto nel Mediterraneo centro-orientale* [*Linear or Hieroglyphic? Comparative Writing Systems in the Central-Eastern Mediterranean*]. Ed. by Marco Mancini and Barbara Turchetta. Rome: Il calamo, pp. 115–142.
- Massias, Nicolas de (1828). *L'influence de l'écriture sur la pensée et sur le langage*. Paris: Firmin Didot.
- Matthews, Peter (1997). *The Concise Oxford Dictionary of Linguistics*. Oxford: Oxford University Press.
- Nerius, Dieter and Gerhard Augst (1988). "Probleme der geschriebenen Sprache". In: *Linguistische Studien*. Vol. 173: *Beiträge zur Schriftlinguistik auf dem XIV. Internationalen Linguistenkongress 1987*. Berlin: Akademie der Wissenschaften der DDR.
- Pei, Mario A. and Frank Gaynor (1954). *A Dictionary of Linguistics*. New York: Rowman & Littlefield.
- Pernigotti, Sergio (1986). "L'antico Egitto [Ancient Egypt]". In: *Sulle tracce della scrittura* [*On the Traces of Writing*]. Ed. by Giorgio Raimondi Cardona. Bologna: Grafis, pp. 25–46.
- Pierides, Demetrios (1875). "On a Digraphic Inscription Found in Larnaca". In: *TSBA* 4, pp. 38–43.
- Rogers, Henry (2005). *Writing Systems. A Linguistic Approach*. Oxford: Blackwell.
- Sallet, Alfred von (1875). "Die Münzen der griechischen Könige von Salamis in Cypern und die denselben zugetheilten moderne Fälschungen". In: *Zeitschrift für Numismatik* 2, pp. 130–137.

-
- Saussure, Ferdinand de (1959). *Course in General Linguistics*. New York: Philosophical Library.
- Trask, Larry (1996). *A Dictionary of Phonetics and Phonology*. London: Routledge.
- Wexler, Paul (1971). "Diglossia, Language Standardization, and Purism: Parameters for Typology of Literary Languages". In: *Lingua: International Review of General Linguistics* 27, pp. 330–354.
- Zima, Petr (1974). "Digraphia: The Case of Hausa". In: *Linguistics: an Interdisciplinary Journal of the Language Sciences* 124, pp. 57–69.

Unicode from a Linguistic Point of View

Yannis Haralambous & Martin Dürst


Abstract. In this paper we describe and comment, from a linguistic point of view, Unicode notions pertaining to writing. After comparing characters with graphemes, glyphs with graphs and basic shapes, character general categories with grapheme classes, and character strings with graphemic sequences, we discuss two issues: the phenomenon of ligatures that stand at the boundary between graphemics and graphetics, and the proposal for the introduction of “QID emojis” which may end up being a turning point in human communication.


1. Introduction

Unicode is a computing industry standard for the encoding, representation, and handling of text. It has been introduced in 1991 and is nowadays practically the only text encoding standard worldwide. Unicode uses architectural principles, but also has to deal with engineering reality, legacy issues, and sometimes even political considerations:

The Unicode Standard is the product of many compromises. It has to strike a balance between uniformity of treatment for similar characters and compatibility with existing practice for characters inherited from legacy encodings. (*The Unicode Standard. Version 12.0—Core Specification* 2019, p. 159)

Unicode is the first encoding in the history of computing that *instructs* the user wishing to write in various writing systems of the world: the Unicode Consortium publishes a 1,018 pages long compendium (*ibid.*)

Yannis Haralambous  0000-0003-1443-6115
IMT Atlantique & LabSTICC UMR CNRS 6285
Brest, France
yannis.haralambous@imt-atlantique.fr

Martin Dürst  0000-0001-7568-0766
Aoyama Gakuin University
College of Science and Engineering
Sagamihara, Japan
duerst@it.aoyama.ac.jp

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1.
Fluxus Editions, Brest, 2019, p. 127–166. <https://doi.org/10.36824/2018-graf-hara1>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

that is comparable, in size and in the amount of details, to the monumental *World's Writing Systems* by Daniels and Bright (1996). And furthermore, the Unicode Consortium has released 14 Standard Annexes, 7 Technical Standards, 6 Technical Reports and 4 Stabilized Reports dealing with issues as manifold as line breaking, bidirectional rendering, vertical text layout, emojis, etc.

In its almost thirty years of existence, the Unicode Consortium has patiently built a technical vocabulary to describe computing issues but also (grapho)linguistic concepts. In this paper, after an introduction to the issue of encoding text, we will consider Unicode's approach to writing, from a linguistic point of view. For this we will adopt a task-oriented approach and will compare two processes: on the one hand, a person reading and understanding text that is displayed on an electronic device, and on the other hand, a machine "reading" and analyzing text from an input flow. Common to these two processes are their extremities: both take Unicode-encoded text as input, both result in "understanding" the text, in the sense of "accessing the various linguistic levels and extracting semantics" contained in it.

The reason for comparing these two—seemingly quite different—processes is that they reveal the double nature of Unicode: in order for humans to read text on screen, Unicode has to supply sufficient information for the text to be adequately rendered (with the help of a rendering engine and information provided in fonts); in order for algorithms to "read" text from an input flow, Unicode has to supply sufficient information to convey all strata of linguistic information. These two needs are complementary and Unicode has been engineered to handle both of them.

This paper is structured as follows: in Section 2 we give a quick review of the fundamentals of grapholinguistics, an introduction to the issues underlying the encoding of text and some extreme cases of text difficult to encode. Section 3 presents the "reading" processes we will refer to in the following sections. Section 4 compares various Unicode terms with grapholinguistic notions. Section 5 deals with ligatures, and Section 6 with emojis.

2. The Context

2.1. Linguistic Fundamentals

When Martinet (1970) defined *double articulation*, he considered phonemes as the lowest level of articulation in the hierarchy of language. A *phoneme* is a distinctive unit in a system and has no meaning per se. It is

through building sequences of phonemes that elementary units of meaning emerge, which we call *morphemes*. This process is called *second articulation*. To verify whether units are distinctive, Martinet used the method of *commutation*: if by replacing a single elementary unit by another in a morpheme, its meaning changes, then these two units are indeed phonemes: “cat” vs. “hat,” “cat” vs. “cut,” “cat” vs. “car”; if not, then they are *allophones* of the same phoneme. As for morphemes, their combination gives access to higher levels of meaning through syntax, and this process is called *first articulation*.

Anis (1988, p. 89) and Günther (1988, p. 77) use the same method to define *graphemes*: in their approach, which is called *autonomistic* because it does not involve phonemics, graphemes are units of written text that have no meaning per se, but are distinctive members of a system. They are defined by commutation, and through their *catenation* (a notion introduced by Sproat (2000, p. 13), see also § 4.4) morphemes emerge: in the English language, <c>, <a>, <t> and <s> have no meaning, but their catenation as <cats> contains two morphemes: <cat>, which represents the concept of “cat,” and <s>, which represents the feature of plural number.

Distinctive units are necessary because humans live in an analog world. As an infinite variety of sounds is perceived by the ear and an infinite variety of shapes is perceived by the eye, our brain has first to select those that participate in linguistic communication (which are called *phones* and *graphs*, respectively), and then to classify them into a finite set of classes by which morphemes are formed. *Graphs* are studied by the discipline of *graphetics* (Meletis, 2015), the name of which is inspired by the analogous discipline *phonetics* which is studying sounds (called *phones*) produced by humans in the frame of oral communication.

Besides *graphs* (the shapes) and *graphemes* (the elementary linguistic units), Rezac (2009) introduces the intermediate notion of *basic shape*, i.e., clusters in the space of possible graphs representing the same grapheme, such as |a|¹ and |ɑ| representing grapheme <a>, or |π| and |ϖ| representing grapheme <π>. Note that these basic shapes do not commute in the context of, e.g., English and Greek language (<cat> ≡ <cat> and <πρός> ≡ <ϖρός>) but can very well commute in other contexts, such as IPA notation system (where <a> and <ɑ> represent different phones), or mathematical notation (where π and ϖ may represent different mathematical variables).

1. In this paper we use the following notation: <.> for graphemes, |.| for graphs, /./ for phonemes and [.] for phones.

2.2. Encoding Text

The Cambridge Dictionary of English defines the verb “to encode” as

to put information into a form in which it can be stored, and which can only be read using special technology or knowledge.

This definition involves three actions: “putting,” “storing,” and “reading”. In our context, “putting” will be understood as “converting analog information into digital form,” or “producing (digital) information on a digital medium,” and we will restrict ourselves to information contained in textual data, where “text” is taken in a rather broad sense, including data in various notation systems such as mathematical formulas, etc. “Storing” can be digital or analog (such as in physically printed material), and “reading” can take different forms, depending on the actor: a human can read an analog text (optically or haptically) produced by mechanical means, or read an analog text produced by a digital device (and hence using digital information), a machine can “read” an analog text (by OCR), or “read” digital data in the sense of a program receiving the data through an input stream.

We will discuss “reading” processes in the next section. In this section we will consider the form in which textual data are converted as a result of the encoding process. Text in natural language (and this is the main target of the Unicode encoding standard) is a complex object with many strata of information. Even if we restrict ourselves to information of linguistic nature, the “encoding” process can be manifold:

1. One of the most common input devices is the computer keyboard, which is functionally a descendent of the typewriter. “Encoding” a text via a typewriter amounts to *keyboarding* it. Keyboarding a text amounts to selecting keys, pushing them and obtaining a 1-dimensional graphetic sequence (Meletis, 2019, pp. 117–120) on the paper. The size of the paper being limited, the typewriter’s carriage return allows the writer to build an 2-dimensional graphetic sequence in areal space. As for the computer keyboard, it also has a carriage-return key, but its use is not necessary since computer memory can be considered as a “page of infinite width” and therefore “encoding” a text through a computer keyboard results in a long 1-dimensional sequence of elementary information units corresponding to keys (or key combinations) pushed by the keyboarder.

The result of this kind of “encoding” is a digital object called *plain text* and this is the type of data Unicode claims to encode.

2. Other legacy text production techniques such as typography have a wider spectrum of visual communication methods (italics, letterspacing, color, etc.) which can be used for various linguistic or paralinguistic functions and therefore can be considered as an integral part

of text and need to be encoded as well. Markup languages such as XHTML (Pemperton et al., 2018) or XSL-FO (Berglund, 2006) handle this kind of encoding efficiently, the result being called *rich text*.

3. Natural language has two main modalities: the written modality and the spoken modality. In languages with shallow orthography such as Italian or fully voweled Standard Arabic, one can easily convert data between these modalities, with little or no information loss; in languages with deep orthography, such as English or Greek, this process requires elaborate algorithms and heavy linguistic resources. By *phonetically annotating* an (encoded) text, one has immediate access to both modalities. A text encoded, e.g., in FoLiA format (Gompel and Reynaert, 2013) can have phonetic and/or phonological annotations.
4. Being a linguistic object, text can be analyzed using traditional linguistic methods, and the results of this analysis can be marked in the text, resulting into what is called *annotated text*. This may seem unnecessary for a human reader who is knowledgeable of the natural language of the text, but can be useful for a human reader learning the language, or for the machine having to process linguistic data. Indeed, the first step of most Natural Language Processing algorithms is a morphosyntactic parse, and obtaining the representation of a text in, e.g., CoNLL-U format (Marneffe et al., 2013) is another kind of “text encoding,” where part-of-speech tags, lemmas and dependency relations are explicitly included.
5. But why stop at the syntactic level? The next step is to perform *semantic annotation* and to encode concepts present in the text and relations between them by aligning them with ontologies, knowledge bases and other semantic resources. This is possible through Semantic Web technologies such as OWL (Bao et al., 2012) and RDF (Hayes and Patel-Schneider, 2014) embedded into the generic markup language XML (Bray et al., 2008). Encoding text in this way allows optimal processing by Natural Language Processing algorithms.

As we see, text “encoding” can be more or less elaborated and rich in information, depending on the target “reader”. When the “reader” is a human, then approaches 1 and 2 are clearly distinct from approaches 3–5. Indeed, the former provide a visual result that can be read by the human, while the latter enrich the text by adding additional information to it—even though one can invent new methods of displaying the additional information, such as interlinear annotations, special GUIs, etc. When the “reader” is the machine, there is no visual stage and the distinction becomes void.

Many large corpora adopt more than one encoding approach. For example, the *Digital Corpus of Sanskrit* (Hellwig, 2010–2019) is a digital object which can be “read” by a human in the traditional way, but also contains full morphological and lexical data. These data can be presented to the human user through a dedicated GUI or can be directly “read” by NLP

algorithms for processing of the text. The *Quranic Arabic Corpus* (Dukes, Atwell, and Habash, 2013) goes even farther and contains dependency syntax and semantic annotation information.

Sometimes the boundaries between the technologies we mentioned become blurry. For example, to state only two examples involving Unicode:

1. In Japanese and Chinese, *ruby* can add the phonetic realization of a morpheme (written in kanji/hanzi characters) using smaller characters from a syllabary (kana in Japanese, bopomofo in Chinese), the latter placed above the former, as in 会社 (“company,” pronounced かいし や *kaisba*). Even though Unicode proclaims that it encodes only plain text, it provides nevertheless three *interlinear annotation characters* for marking the begin of the base sequence, the separation between base and annotation and the end of the annotation sequence. XHTML also provides markup for ruby (the ruby element) and this is the method recommended by the W3C (see Sawicki et al. 2001, as well as Dürst and Freytag 2000). Ruby, as an annotation, is essentially phonological and morphological since, traditionally, ruby bases are morphemes and not individual characters—it therefore overlaps approaches 3 and 4.
2. There exist Unicode characters with empty visual representation. These characters carry information of morphological, syntactic or semantic nature, e.g., the SOFT HYPHEN character marks boundaries of graphical syllables (and is useful for obtaining correct hyphenation); the INVISIBLE SEPARATOR character marks consecutive symbols as being part of a list (a property of syntactic nature); and the INVISIBLE TIMES character marks consecutive symbols as being multiplied (a binary algebraic operation with very precise semantics). The information carried by the INVISIBLE SEPARATOR and INVISIBLE TIMES characters can also be represented by markup in a markup language such as MathML (Carlisle, Ion, and Miner, 2014): the apply and times elements.

The standard way of producing written text, as described in Meletis (2019, pp. 117–120), is to catenate graphs into *1-dimensional graphetic sequences* to fill *linear space*, until reaching the “page” boundary and then continuing on the next line in order to fill *areal space*. This approach, which is the standard approach of legacy typography, is used by Unicode and rendering engines. It inherently assumes that the geometry of 1-dimensional graphetic sequences—as long as there is no 2-dimensional higher structure such as a list or a table—carries no syntactic or semantic information.

There are cases where human creativity has transcended this model, and we will present three examples (cf. Fig. 1). It is legitimate to raise the

question whether these cases can be “encoded” by the machine without information loss. They are the following:

- (a) a page from Mallarmé’s “Un coup de dés jamais n’abolira le hasard” (“A throw of the dice will never abolish chance”) where the reading process is spatially and temporally structured by horizontal and vertical gaps, font size, font style and uppercasing. To achieve the visual result with the appropriate precision while keeping access to textual content, XHTML and XSL-FO are not sufficient and a markup language for describing two-dimensional vector and mixed vector/raster graphics, such as SVG (Bellamy-Royds et al., 2018) is necessary;
- (b) Apollinaire’s *calligram* “La colombe poignardée et le jet d’eau” (“The stabbed dove and the fountain”), where not only graphemes form a drawing, but text meaning and image are in constant interaction: For example, the soft and immaculate character of the dove’s wings is strengthened by the text fragments on their contours: “douces figures” (“soft figures”) and “lèvres fleuries” (“flourished lips”) and by six female given names followed by the question “où êtes-vous ô jeunes filles” (“where are you oh young girls?”). Also, the wound of the stabbed dove is drawn with the words “et toi” (“and you”). Here again a markup language such as SVG is necessary in order to place graphemes on curved paths while maintaining the linearity of the poem’s text, and to encode the correspondences between parts of the drawing and text segments. One can even envision an abstract hierarchical description of the drawing (involving the dove, its head, wings and queue, wings being made of feathers, etc.) where each element is linked to a text segment, so that the poem inherits the graphical structure of the drawing and so that we have an alignment between linguistic and pictorial hierarchical structures (see Fig. 2 for a small excerpt of such a structure). No less than that would be necessary to capture the subtle interaction between image and text;
- (c) and finally a Kufic calligraphy of a Quranic verse: ﴿لَا يُكَلِّفُ اللَّهُ نَفْسًا إِلَّا وُسْعَهَا﴾ (“God does not burden any soul beyond its capacity,” 2:286), written as a spiral starting from the lower right corner and going clockwise inwards so that the last word is in the middle of the drawing. Here, the act of recognizing the text inside the labyrinthian drawing symbolizes, in the frame of Muslim religion, the discovery of the word of God in the world, which is His book.

For such a calligraphy to have a dual text/image nature, a markup language such as SVG is again necessary, but also an ad hoc font with glyphs dynamically drawn out of generic “metagraphs” for the

LE NOMBRE

EXISTAT-IL
 autrement qu'hallucination éparsée d'agonie
 COMMENÇAT-IL ET CESSAT-IL
 sourdant que nié et clos quand apparut
 enfin
 par quelque profusion répandue en rareté
 SE CHIFFRAT-IL
 évidence de la somme pour peu qu'une
 ILLUMINAT-IL

LE HASARD

Choit
 la plume
 rythmique suspens du sinistre
 s'ensevelir
 aux écumes originelles
 naguères d'où sursauta son délire jusqu'à une cime
 flétrie
 par la neutralité identique du gouffre

425

(a)

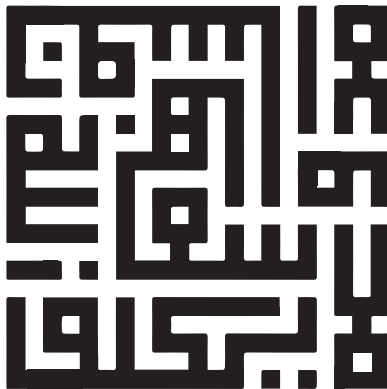
LA COLOMBE POIGNARDÉE ET LE JET D'EAU

Douces figures poi^{rdée}
 MIA Chères lèvres fleuries
 YETTE MAREYE
 ANNIE et toi LORIE
 où vous MARIE
 jeunes filles
 MAIS
 près d'un
 jet d'eau qui
 pleure et qui prie
 cette colombe s'extasie

Tous les souvenirs de nag^{uères} ? O^ù sont Reynal Billy Dalize
 O mes amis partis en guerre Dont les noms se mélancolisent
 Jaillissent vers le firmament Comme des pas dans une église
 Et vos regards en l'eau dormant Où est Crémnitz qui s'engagea
 Meurent mélancoliquement Peut-être sont-ils mort déjà
 Où sont-ils Braque et Max Jacob De souvenirs mon âme est pleine
 Derain aux yeux gris comme l'aube e^t le jet d'eau pleure sur ma peine

CEUX QUI ONT PARTI À LA GUERRE AU NORD SE BATTENT MAINTENANT
 Le soir tombe O sanglante mer
 Jardins où saigne abondamment le laurier rose fleur guerrière

(b)



(c)

FIGURE 1. Three examples where the size, style, position and form of graphetic sequences participate in meaning production

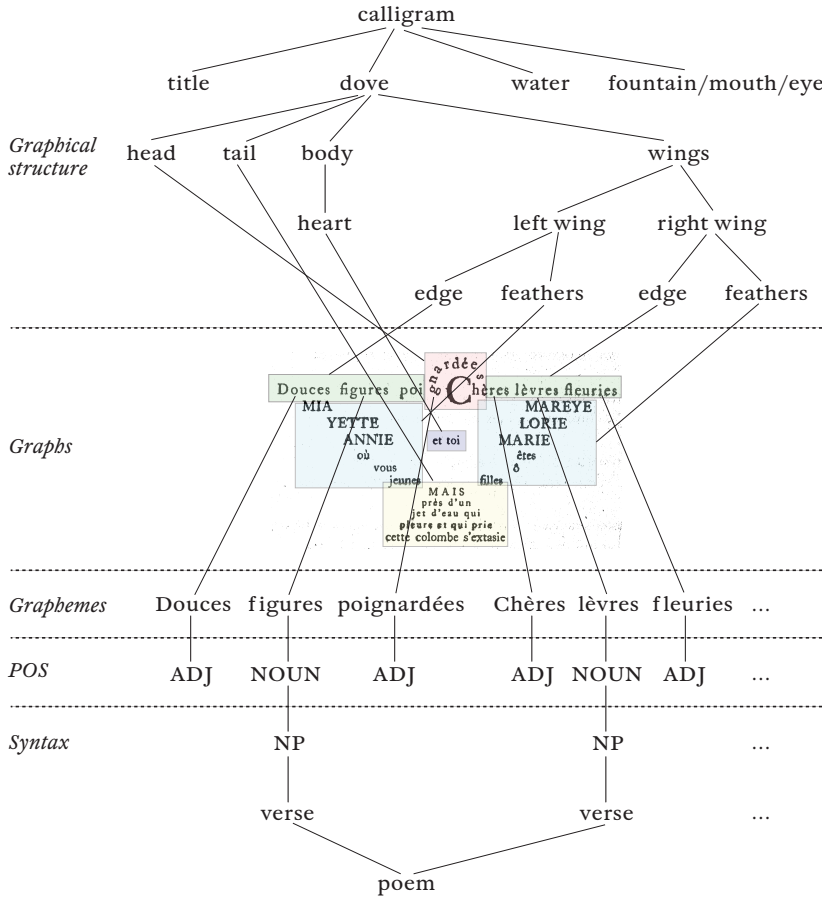


FIGURE 2. Levels of analysis of a small fragment of Apollinaire’s “La colombe poignardée et le jet d’eau”

specific textual utterance (André and Borghi, 1990; Bayar and Sami, 2010).

The aim of this section was to introduce the reader to the problematics of “text encoding,” to the various technologies available, to their mutual boundaries and overlaps, and to their limits illustrated by three examples of texts, the visual methods of which exceed the common meso/macrographetic model².

2. These terms have been introduced in Meletis (2015) and were inspired by the terms “meso-” and “macrotypography,” introduced by Stöckl (2004, p. 22).

In the remainder of this paper we will adopt a task-oriented approach and examine Unicode in the frame of three “reading” processes, differing by their actors: human or machine.

3. Three “Reading” Processes

The field of *perceptual graphetics* (Meletis, 2015) deals with the influence of the materiality of writing on perception, recognition and reading, and there has been abundant research on the particular case of perceptual graphetics of electronic devices (computer screen, tablets, smartphones, etc.). In this research domain, the displayed text is considered as a starting point and the objects of study are mainly the human perception of the signal emitted by the machine and the cognitive processes involved in recognition and understanding of textual data.

We will extend the “reading” process to the situation where the emitter and receiver are machines, and the channel is purely digital (so that no optical intermediation is involved). We therefore describe three situations where text is “read,” corresponding to three different processes, illustrated by Fig. 3 where the text consists of the morpheme “cat”:

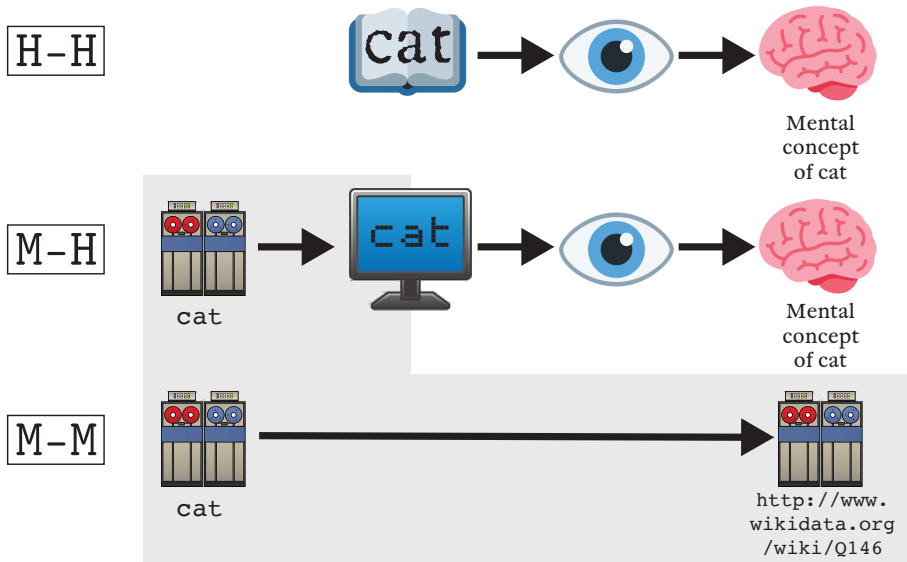


FIGURE 3. Processes $\boxed{\text{H-H}}$, $\boxed{\text{M-H}}$ and $\boxed{\text{M-M}}$ involving the morpheme <cat> and the concept of cat. Grey background denotes the digital world.

- H-H (“Human → Human”) is the process of reading from paper (either manuscript or printed): The eye sees graphs in the material form of ink on a the paper surface, the brain recognizes graphemes, combines them into morphemes and accesses the mental concepts they represent;
- M-H (“Machine → Human”) is the same process, but this time the reading surface is a computer monitor or other electronic device. The computer has the word *cat* stored, encoded in Unicode, and transmits this information to a rendering engine which extracts data from a font, builds an image, and transmits this image to a display device. At the boundary between the digital (grey background in the figure) and the analog world, the device receives the data and displays the word on its surface. The rest of M-H is the same as H-H;
- M-M (“Machine → Machine”) is the process of accessing the same concept through Natural Language Processing algorithms. The input is the same: the word *cat* encoded in Unicode. NLP algorithms have to (a) detect the current language; (b) use this information to detect words/morphemes; (c) use the context and linguistic resources to find POS tags; (d) disambiguate: is “cat” the felid? or a jazz enthusiast? or the pointed piece of wood that is struck in the game of tipcat? or is it a sturdy merchant sailing vessel?³ Once disambiguated, the intended (or most probable) concept is represented by an IRI (*Internationalized Resource Identifier*, see Dürst and Suignard (2005)) pointing to an item of the knowledge base Wikidata, namely <http://www.wikidata.org/wiki/Q146>. This item represents the concept of “cat” as a felid.⁴

The diagram of Fig. 3 obviously oversimplifies the real processes—its purpose is to show the *dual purpose* of Unicode, which has to provide sufficient information to a rendering engine to be able to correctly render graphemes on the display so that process M-H can succeed, but also to provide sufficient information to NLP algorithms for the process M-M to succeed as well.

In process M-M, Natural Language Processing algorithms need to have sufficient information to access all facets of the input considered as a linguistic object. Unicode has been engineered to allow this kind of process. But it also allows process M-H: in this case, the system transmits the string of Unicode characters to a rendering engine, which will load a *font* that maps characters to images (called *glyphs*) displayed to the reader (cf. Haralambous 2007).

3. The various meanings of the word “cat” listed here are taken from the English Wiktionary entry <https://en.wiktionary.org/wiki/cat>.

4. We have chosen Wikidata as an example, but other knowledge bases also exist, such as WordNet or Yago. See §6 for additional information on Wikidata.

Usually the choice of the font involved in the rendering process depends on the knowledge domain (computer science and mathematics publications are often typeset in Computer Modern, most other scientific disciplines in Times) or on the publisher’s graphical signature, and therefore its contribution to meaning production is secondary and mostly connotative. For that reason Unicode does not encode fonts and this information has to be added using higher protocols such as markup languages⁵ or stylesheets. Once again there are cases where human creativity has transcended this convention and has raised the choice of font to the status of important factor in the production of meaning. The reader can see an example in Fig. 4, an Austrian publicity: “Gehen Sie wählen! Andere tun es auch.” (“Go vote! Others do it too.”), where the change to a broken script in the second sentence narrows the referent of the noun “others” to right-wing extremists⁶ and thereby denotes political orientation.

For the time being, cases involving font choice or specific geometric arrangements of text can successfully go through the $\boxed{\text{M-H}}$ process but, to the authors’ knowledge, no NLP algorithm is yet capable of extracting the meaning produced by geometric and graph(et)ic features, mainly because NLP relies on the “plain text” model and discards any information of graphetic nature.

This leads us to the two main questions of this text:

1. How does Unicode model writing, in order to handle both the $\boxed{\text{M-H}}$ process and the $\boxed{\text{M-M}}$ process?
2. What are the fundamental notions of Unicode, and how do they relate to linguistic notions and processes?

In the following sections we will explore the Unicode notions of character, character category, glyph, and character string, and relate them



FIGURE 4. Austrian publicity: “Go vote! Others do it too.” (Also in Dürscheid 2016, p. 232 and Schopp 2008.)

5. The SVG (Bellamy-Royds et al., 2018) markup-language not only allows choice of font by name, but also provides XML markup for designing entire fonts which can be stored internally in the document or be used remotely.

6. Which is actually ironic, since it was Hitler who prohibited the use of broken scripts in 1943, cf. Haralambous (1991).

to the linguistic notions of grapheme, grapheme class, graph and basic shape, 1-dimensional graphetic sequence, etc.

4. Unicode

4.1. Characters vs. Graphemes

The atomic unit of Unicode is the *character*. This term has its roots in the encodings of the stone age of computing (FIELDATA in 1960, ASCII in 1963, see Mackenzie 1980 and Haralambous 2007). In the early sixties, a “character” was a “specific bit pattern⁷ and an assigned meaning”. The “meaning” was either a “control meaning” (ringing a bell, delete the previous character, etc.) or a “graphic meaning”. A “graphic meaning” was either an “alphabetic,” or a “numeric,” or a “special” (i.e., punctuation and logographs such as %, #, @, etc.). “Alphabets” were defined as “letters in the alphabet of a country” (Mackenzie, 1980, p. 16). This naive and Eurocentric approach is due to the limited use of text in computers of that period.

Unicode being a descendent of ASCII, it inherited the “character” term and introduced a panacea of additional technical terms, some of which we will try to consider in the following, from a linguistic point of view.

Probably to avoid conflict with the ancestral ASCII standard, the term “character” per se is never defined in the Unicode specification. Defined are four specializations of this term: “abstract character,” “encoded character,” “deprecated character,” and “noncharacter”.

According to the Unicode specification, an *abstract character* is defined as

a unit of information used for the organization, control, or representation of textual data. (D7 in U§3.4⁸)

We can’t help notice that this definition takes the notion of “textual data” and its perimeter for granted. This comes as no surprise since the notion of “text” is polysemic and depends on the disciplinary context. In the *Cambridge Dictionary of Linguistics* (Brown and Miller, 2013), “text” is defined as follows:

7. On the lowest level of computer memory, *bits* are binary values b_i , their concatenation $b_n b_{n-1} \dots b_1 b_0$ allows the representation of integer numbers through the formula $\sum_{i=0}^n 2^{b_i}$. A “bit pattern” is a sequence of bit values, for example 01100001 corresponds to number 97.

8. In the following we will denote, for the sake of brevity, a section ** from the Unicode Standard Version 12.0 (*The Unicode Standard. Version 12.0—Core Specification* 2019) by “U§**”.

The term originally denoted any coherent sequence of written sentences with a structure, typically marked by various cohesive devices. It has been extended to cover coherent stretches of speech.

But then again, in the same dictionary, there are no entries for “written sentence” and for “writing,” used in this definition. A “sentence” is defined as

the largest unit handled by grammar,

and “grammar” is defined as either (in the narrow sense) the “morphology and syntax of a language,” or (in the broad sense) the “morphology, syntax, phonology, semantics and even the pragmatics of a language”.

A more general definition of “text” is given in Wikipedia:

In literary theory, a *text* is any object that can be “read,” whether this object is a work of literature, a street sign, an arrangement of buildings on a city block, or styles of clothing. It is a coherent set of signs that transmits some kind of informative message,

followed by a reference to Lotman (1977). If we apply this definition to processes $\boxed{\text{M-H}}$ and $\boxed{\text{M-M}}$ we come to the fact that the purpose of text is to be “read,” be it by a human or by a machine. “Reading,” in our case, comes down to:

1. detecting and identifying the text’s elementary units,
2. applying second articulation to extract morphemes from their combination, and then
3. applying first articulation to obtain meaning from the combination of morphemes.

Our statement is that these three operations apply not only to process $\boxed{\text{M-H}}$, but also to process $\boxed{\text{M-M}}$. To start, the machine is informed by various mechanisms that it is “reading” Unicode characters and not, for example, pixel data. It is also informed on the way of reading these data, in order to convert them appropriately to elementary text units⁹. Once the machine is aware of the fact it is reading Unicode characters, identifying them becomes possible through the notion of “encoded character”:

An *encoded character* is an association (or mapping) between an abstract character and a code point, (U§3.4)

where a *code point* is defined as follows:

9. The details on the various ways of storing elementary text units on the machine level, that is using bits and bytes, have no incidence on the linguistic study of Unicode. The interested reader is invited to read U§2.5 and U§2.6.

A *code point* is any value in the Unicode codespace (U§3.4)

and the *Unicode codespace* is

the range of integers $\{0, \dots, 1,114,111\}$. (U§3.4)

In other words, in the frame of process $\boxed{\text{M-M}}$, step 1 of the “reading” process, namely *identification of elementary units* (called “encoded characters”) is *trivial* for the machine, since they are represented in memory by unique numbers¹⁰. No extra effort is required.

Not so for the human in the frame of process $\boxed{\text{M-H}}$, where elementary units are distinctive elements of a system and, as such, must be recognized by the readers, provided they are knowledgeable of the corresponding writing system or notation system. Here, the “elementary unit” (of text) corresponds to the linguistic notion of *grapheme*, as defined by Anis (1988) and Günther (1988).

So how do the notions of “character” (abstract or encoded) and “grapheme” compare?

The Unicode Consortium avoids taking position in favor or against the autonomistic approach and avoids using the term “grapheme”. Indeed, in the Unicode specification it is stated that

A character does not necessarily correspond to what a user thinks of as a “character” and should not be confused with a *grapheme*. (U§3.4)

As a critique to this statement we argue that:

- in the frame of process $\boxed{\text{M-H}}$, a character is *exactly* the information that, after being channeled through rendering engines, allows the human to recognize a grapheme without ambiguity, so it functionally corresponds to a grapheme;
- in the frame of process $\boxed{\text{M-M}}$, a character is exactly the information that is necessary to the machine to perform natural language processing, similarly to graphemes that are the information necessary to the human to process language,

and therefore one can conclude that, functionally, the notions of character and grapheme are quite close.

Nevertheless characters do not always represent graphemes. The main discrepancy between the two notions is due to the fact that some scripts (such as the Latin script or the Cyrillic script) are used for more

10. In some sense, encoded characters can be considered as signs, code points being their signifiers and abstract characters their signifieds. As with Saussurean signs, the relationship between signifier and signified is arbitrary, in the sense that there is no reason why the abstract character LATIN LETTER A is represented by code point 97.

than one language, and Unicode targets *all* languages simultaneously. To take two examples, according to Grzybek and Rusko (2009, p. 33), there is a <ch> grapheme in the Slovak language, and according to Wmffre (2008, p. 598), there is a <c'h> grapheme in the Breton language—none of these is a Unicode character. This comes from the fact that Unicode is bound to follow conventions such as national encodings or keyboard layouts, and there has never been a Slovak encoding or keyboard featuring <ch> or a Breton encoding or keyboard featuring <c'h>.

There are even Unicode characters that do not correspond to graphemes in *any* language, such as invisible Unicode characters (various spaces, SOFT HYPHEN, etc.), pictographs, emojis, etc. Also there is a large amount of characters that are graphemes in one language but not in another language using the same script (such as <č> used in Czech but not in German). More interesting is the case of characters which are graphemes in language A, are not graphemes in language B, but can still be used in B as allographs of some grapheme. For example, the letters <ڪ> and <ك> which commute in Sindhi: <ڪنڊ> (“ear”) ≠ <ڪنڊ> (“sugar”); in Arabic and other Arabic-script languages they are just allographs of the Arabic <ك>, and hence do not commute: <ڪتاب> ≡ <ڪتاب> (both “book” in Arabic).



FIGURE 5. French logos using allographs that are graphemes in other languages

This phenomenon is easier to observe in graphic design, where the use of allographs is a creative design method. In the French-language examples of Fig. 5 one can observe the use of allographs <İ>, <Ē>, <Ā>, and <í>, instead of <I>, <É>, <Ā> and <i>; these allographs are graphemes in other languages (Turkish for <İ>, Latvian for <Ē> and <Ā>, Italian, Spanish and elsewhere for <í>) and therefore are accessible to the designer through Unicode-encoded fonts.

4.2. Glyphs vs. Graphs and Basic Shapes

Meletis (2015, p. 117) defines *graphs* for the German language as follows:

Die kleinste, nicht weiter durch Leerstellen getrennte Einheit ist der Graph (dies gilt insbesondere für Druckschrift und nur eingeschränkt für

Handschrift), der jeweils einen einzigen segmentalen Raum ausfüllt und im alphabetischen Schrifttyp durch Buchstaben, aber auch nicht-alphabetische Graphen wie Interpunktions- und Sonderzeichen sowie Ziffern verkörpert wird.¹¹

This definition can easily be applied to scripts with separated atomic units (alphabetic scripts, South-East Asian scripts, Chinese script). In the case of scripts systematically connecting atomic units (Arabic and Syriac scripts, Devanagari, etc.) the condition of “smallest entity not separated by blank spaces” cannot be applied. In the case of printed text, one can refer to the historical segmentation of the connected script into types to obtain a (not perfect but reasonable) solution to the problem how to segment a shape into its constituent graphs. In the case of handwriting there is no clear segmentation and fuzzy logic has to be applied to the mapping of each part of the drawing to individual graphs contained in it.

The advantage of this definition of *graph* is that it doesn't presuppose knowledge of graphemes and of higher linguistic units: one can take a printed text in an unknown language, subdivide it into elementary graphical units and move on to the next graphetic levels (1-dimensional graphetic sequences in linear space, 2-dimensional graphetic sequences in areal space, page).

In the Unicode Standard there is no proper definition of *glyph*. The closest we can get to obtain a definition would be through the following sentence:

Glyphs represent the shapes that characters can have when they are rendered or displayed. (U§2.2)

What is understood in this definition is that the glyphs must be rendered or displayed in such a way that the characters they represent can be visually recognized by readers knowledgeable of at least one language in which the characters are used.¹²

All occurrences of the term “glyph” in the Unicode Standard refer to shapes obtained by rendering characters. This excludes shapes drawn by hand in a drawing application, and any text obtained by a means different than Unicode characters rendered by a rendering engine.

11. “The graph is the smallest entity [of the model] that is not separated by blank spaces (this is valid predominantly for printed text and only partially for handwritten text) and that occupies a single segmental space. In alphabetic writing systems, graphs are materialized by letters but also by non-alphabetic signs such as punctuation and special signs and digits.”

12. Technically the rendering of an arbitrary Unicode character is provided by rendering engines, which use data from fonts, and in font technologies there is absolutely no restriction on the shape that can be used for a specific Unicode character: one can easily create a font rendering the character LATIN CAPITAL LETTER A by the glyph |B|.

So, if we stick to this excerpt of the Unicode Standard, then (a) glyphs are an aspect of characters for which Unicode takes absolutely no responsibility, (b) absolute freedom is granted to font vendors to render characters as they like and (c) only the degree of commercial success of a font can determine the legitimacy of its glyphs as representatives of given characters. This may seem an overstatement for the common scripts, but becomes a real problem for rare scripts for which only very few fonts exist: Dürscheid (2018, §4) qualifies the Unicode Consortium as a “gatekeeper” of characters, in a similar way the font industry becomes the “gatekeeper” of glyphs of rare scripts¹³.

Fortunately Unicode avoids this anarchy situation by introducing an additional notion: the one of *representative glyph*:

The identity of a character is established by its character name and *representative glyph* in the code charts.

A character may have a broader range of use than the most literal interpretation of its name might indicate; the coded representation, name, and representative glyph need to be assessed in context when establishing the identity of a character. For example, FULL STOP can represent a sentence period, an abbreviation period, a decimal number separator in English, a thousands number separator in German, and so on. The character name itself is unique, but may be misleading.

Consistency with the representative glyph does not require that the images be identical or even graphically similar; rather, it means that both images are *generally recognized to be representations of the same character*. Representing the character LATIN SMALL LETTER A by the glyph “X” would violate its character identity. (U§3.3, emphasis introduced by us)

Representative glyphs for all Unicode characters can be found in the Unicode Code Charts¹⁴. The notion of representative glyph is very interesting because

1. it reveals the insufficiency of the intensional description of characters;
2. it induces an operational definition of glyphs: *a glyph is a shape that is generally recognized to be a representation of a character*. This definition still involves Unicode characters, but not the rendering process anymore;
3. it shows that the relation between characters and glyphs has a socio-linguistic facet: a glyph represents a given character if and only if there is a *community of people* recognizing it as such.

The notions of *graph* in linguistics and *glyph* in Unicode may intuitively seem equivalent, but the ways they are defined makes them difficult to

13. In the sense that a Unicode user without the necessary competency for creating a font with the appropriate glyphs is forced to use glyphs provided in existing fonts.

14. <https://www.unicode.org/charts/>.

compare: *graphs* are defined as units in a graphetic system, while *glyphs* are defined as socially recognizable renderings of a given character.

It is interesting to note that in Unicode, kanji/hanzi characters have not one but as many as six representative glyphs, corresponding to graphs used in China, Hong Kong, Taiwan, Japan, Korea and Vietnam. For example, the character 伶 with codepoint 4F36₁₆ (“clever,” “actor”) is presented in the following way in the Unicode code chart:

As the reader can see there are three shape families: (a) the first and sixth graphs (China and Vietnam) have a drop-like *diǎn* stroke |丿| under the “roof”; (b) the second and third graphs (Hong Kong and Taiwan) have a straight horizontal *béng* stroke |一| under the “roof”; (c) the fourth and fifth graphs (Japan and Korea) have a different lower-right component, consisting of a *béng-zhè-gōu* stroke |𠃉| and a *shù* stroke |丨| (see Haralambous 2007, pp. 154–155 and Myers 2019, pp. 13–14).

We claim that these three graphs belong in fact to three different *basic shapes*, in the sense of Rezec (2009). As they are obtained by the use of different fundamental strokes, the graphs will necessarily remain different in all possible realizations belonging to three disjoint clusters. There will never be “intermediate” cases since the fundamental strokes have to be recognizable by design as distinctive parts of the system.

4.3. Character General Categories vs. Grapheme Classes

Let us now turn to issues of classification. In the usual classification of graphemes into *logograms* and *phonograms*, the latter are defined by their relation to speech. Dürscheid (2016, p. 74) defines phonograms as follows:

Phonogramme (= Lautzeichen) sind Zeichen, die ausschließlich auf die lautliche Ebene des Sprachsystems bezogen sind.¹⁵

Such a definition is not compatible with Anis’s autonomistic approach, which considers writing without any a priori relation to speech. Anis (1988) divides graphemes into three classes, namely *alphagrams*, *topograms* and *logograms*. His definition of an *alphagram* is as follows:

ces unités distinctives, dénuées de sens par elles-mêmes, sont les composantes des unités significatives. Comme les phonèmes, les alphagrammes relèvent de la *seconde articulation*.¹⁶

15. “*Phonograms* (= Signs for sound) are signs that refer exclusively to the oral level of the language system.”

16. “These distinctive units, meaningless per se, are components of significative units. Like phonemes, alphagrams are part of *second articulation*.”

A *topogram* (Anis, 1988, p. 116) is essentially punctuation: topograms contribute to the structure and segmentation of sequences of alphagrams and logograms. As for *logograms*, they are global units having a signified (ibid., p. 139)¹⁷.

Typical examples of alphagrams are members of alphabets, of abjads, of abugidas, of syllabaries. Typical examples of logograms are graphemes such as <&>, <§>, currency signs <\$>, <€>, etc., mathematical symbols <5>, <∇>, etc., general symbols <☉>, <σ>, <ℵ>, etc.

Unicode provides a similar classification of characters in the form of a mandatory normative¹⁸ property of characters, called *general category* (U§4.5). This classification is quite different from the linguistic one:

1. all alphagrams belong to general category “L” (for “Letter”), with the following subcases: “Lu” (“uppercase”), “Ll” (“lowercase”), “Lt” (“titlecase”), “Lm” (“modifier”) or “Lo” (“other”). The general category “L” is the most populated in Unicode: it amounts to 89.84% of the total set of characters. Among them, 96.13% are caseless and therefore belong to category “Lo” (caseless alphabets, abjads, abugidas and syllabaries, and most importantly, all Chinese characters);
2. many logograms¹⁹ such as <&>, <@>, <%>, etc., are of Unicode general category “Po” (“punctuation, other”), a contradiction to their linguistic classification;
3. in the case of mathematical symbols, Unicode uses two general categories: “N*” for numbers, and “Sm” for other mathematical symbols, such as <+>, <≤>, etc. The “N*” general category contains the subcategories “Nd” (decimal digits), “No” (fractions, numbers larger than 9, circled or parenthesized numbers) and “Ni” (Roman, Hangzhou, Bamum, Greek acrophonic, Gothic, Old Persian and Cuneiform numerals);
4. The general category “So” (“symbol, other”) is a catch-all. It includes symbols such as <©>, <°>, <⊕>, but also emojis, musical symbols, technical drawing symbols, circled or parenthesized letters/ideograms/syllables, box drawing symbols, Braille patterns, Chinese radicals, Chinese fundamental strokes, hexagrams, Phaistos disk signs, sign-writing gestures, Mahjong and domino tiles, playing cards, alchemical symbols, as well as the single (!) character ARABIC LIGATURE BISMILLAH AR-RAHMAN ARRAHEEM

17. Anis does not distinguish between the iconic and the indexical semiotic function and therefore considers pictograms as being a special case of logograms. We will not adopt his choice and will consider pictograms as being distinct from logograms, even though the distinction can sometimes be blurry.

18. In the sense that Unicode-compliant software has to respect it.

19. Other than Chinese characters, which are not pure logograms since they can have different amounts of semanticity and phoneticity, cf. Haralambous (2013). As already mentioned, Chinese characters belong to category “L” (“letters”).



representing the Arabic sentence “In the name of Allah, the Beneficent, the Merciful”. Category “So” characters represent 5% of the total number of characters.

The linguistic classification of graphemes and the Unicode classification of characters differ in their finalities:

- the former focuses on the way graphemes contribute to meaning extraction: alphagrams are part of second articulation, and hence meaning emerges from their catenation; topograms structure grapheme sequences, and hence serve on the syntactic level; logograms represent morphemes;
- the latter focuses on the way characters are used by software: characters that serve in linguistic processes (“letter” category) are separated from punctuation, from mathematical symbols, and from symbols in general (among which numerous emojis). General categories are used in texts such as the *Unicode Standard Annex on Text Segmentation* (Davis, 2019b), which defines the boundaries of a “word” or of a “sentence” using general categories, or the *Unicode Standard Annex on Line Breaking* (Heninger, 2019), which gives guidelines to line breaking algorithms, based on general categories.

4.4. Character Strings vs. Grapheme Sequences

A text rarely consists of a single grapheme²⁰. Most often humans produce *sequences* of graphemes. Contrary to phonemic input which is linear due to the structure of human speech production organs, grapheme sequences are usually materialized on a 2-dimensional surface. The linear order of phonemes is often represented by a similarly linear order of graphemes (like the ones the reader is reading at this moment), but there are exceptions. A nice example is the Khmer script: the sequence of graphemes representing phonemes /kk/ is written <ក្ក> and when one adds an additional grapheme representing the /r/ phoneme, the sequence representing the phonemic sequence /kkr/ is written as <ក្ករ> (the <រ> grapheme is written to the left of the previous ones), and if one adds a vowel /iə/ this will surround the preceding graphemes: <ក្ករី> (example taken from Haralambous 1994b).

In linguistics, grapheme sequences have been studied by Sproat (2000), in the frame of generative phonology theory introduced by

20. As always, there are exceptions to this rule, such as the title of the recently published book 心 (Kazuo, 2019).

Chomsky and Halle (1968). In this theory one admits the existence of two levels of representation of phonological data: the *underlying form* and the *surface form* (with the possibility of any number of intermediate levels). The latter is obtained by sequentially applying phonological rules to the data of the former (every intermediate level being the output of some rule). A sequence of rule applications going from underlying to surface level is called a *derivation*. Sproat (2000) states that *graphemes can be obtained by using derivations from the same underlying representation as phonemes*, i.e., the graphemic surface representation can be obtained by derivations of the same underlying representation used to obtain surface phonemes. Sproat furthermore claims that this derivation is a *regular relation* in the sense of finite state transducers (Kaplan and Kay, 1994), and that it is consistent throughout the vocabulary of a given language.

Regular relations are context-free, therefore if γ is a derivation and $a \cdot b$ is the catenation of two underlying representation units, then $\gamma(a \cdot b) = \gamma(a) \cdot \gamma(b)$. In fact, according to Sproat (2000), we have not a single but five *catenation operators*, namely $\vec{\cdot}$, $\overleftarrow{\cdot}$, \downarrow , \uparrow , and \odot , representing placement of the second grapheme on the right, on the left, underneath, on top of, or around the first grapheme.

For example, the derivation rules for Korean hangul are as follows (ibid., p. 43):

1. for syllables σ_1 and σ_2 , $\gamma(\sigma_1 \cdot \sigma_2) := \gamma(\sigma_1) \vec{\gamma}(\sigma_2)$;
2. for onset-nucleus ωv and coda κ , $\gamma(\omega v \cdot \kappa) := \gamma(\omega v) \downarrow \gamma(\kappa)$;
3. when the coda κ is complex: $\kappa = \kappa_1 \cdot \kappa_2$, then $\gamma(\kappa_1 \cdot \kappa_2) := \gamma(\kappa_1) \vec{\gamma}(\kappa_2)$;
4. for onset ω and nucleus v , either
 - (a) $\gamma(\omega \cdot v) := \gamma(\omega) \vec{\gamma}(v)$, when v belongs to the vertical jamo class, or
 - (b) $\gamma(\omega \cdot v) := \gamma(\omega) \downarrow \gamma(v)$, when v belongs to the horizontal jamo class;
5. (rule added by us) sometimes the nucleus v is complex and hence can be written as $v = v_1 \cdot v_2$ where v_1 is horizontal and v_2 is vertical, then we first apply rule 4(a) to $\omega v_1 \cdot v_2$ and then rule 4(b) to $\omega \cdot v_1$.

As an illustration, let us apply these rules to Hangul syllable $\langle \text{ㅏ} \rangle$: it consists of an onset $\langle \text{ㄱ} \rangle$, a nucleus containing two jamos $\langle \text{ㅏ} \rangle$ and $\langle \text{ㅣ} \rangle$ of which the first is horizontal and the second vertical, and a coda consisting of two jamos $\langle \text{ㄹ} \rangle$ and $\langle \text{ㅇ} \rangle$. According to rule 5, we first apply rule 4(a) to $\langle \text{ㄱ} \text{ㅏ} \rangle \cdot \langle \text{ㅣ} \rangle$ to get $[[\langle \text{ㄱ} \text{ㅏ} \rangle] \vec{\langle \text{ㅣ} \rangle}]$ and then rule 4(b) to $\langle \text{ㄱ} \rangle \cdot \langle \text{ㅏ} \rangle$ inside it, to obtain $[[\langle \text{ㄱ} \rangle \downarrow \langle \text{ㅏ} \rangle] \vec{\langle \text{ㅣ} \rangle}]$. Then we apply rule 3 to the coda $\langle \text{ㄹ} \rangle \cdot \langle \text{ㅇ} \rangle$ to obtain $[\langle \text{ㄹ} \rangle \vec{\langle \text{ㅇ} \rangle}]$, and finally rule 2 to join onset-nucleus and coda, in order to obtain

$$[[\langle \text{ㄱ} \rangle \downarrow \langle \text{ㅏ} \rangle] \vec{\langle \text{ㅣ} \rangle}] \downarrow [\langle \text{ㄹ} \rangle \vec{\langle \text{ㅇ} \rangle}]$$

as decomposition of $\langle \text{ㅏ} \rangle$. Sproat calls this kind of formal grammar, a *planar regular grammar*.

Among the various applications of these rules there is also diacritization: the grapheme $\langle \hat{a} \rangle$ can be represented by $\langle \text{a} \rangle \uparrow \langle \hat{\ } \rangle$.

It is noteworthy that planar catenators can be applied on all graphic levels: inside a grapheme, between grapheme and diacritic, between graphemes to form morphemes, between morphemes to form lines of text, between lines of text to form paragraphs and pages, between pages to form books, similarly to the graphetic model of Meletis (2015).

In Unicode, there are two notions corresponding to the linguistic notion of grapheme sequence:

1. *combining character sequences*, where we deal with a single “base” character and one or more diacritics, and
2. *character strings*, where we deal with more than one base character.

In the first case, we use the operation of *combination*: a character, which has to be of category other than “M” (“combining mark”) is followed by one or more *combining characters*, i.e., characters of category “M”. For example, to obtain the rendering |â| one can use two characters: LATIN LETTER A followed by COMBINING CIRCUMFLEX ACCENT. In other words: when rendering this sequence of Unicode characters, Unicode-compliant software has to place the glyph of the circumflex accent upon the glyph of the character preceding it.



Combination is a very powerful feature because one can combine any sequence of combining characters (there are 2,268 of them) with any of the 121,490 graphic characters, which results in an astronomical number of combinations. *Interscript* combination is not very frequent but it may happen, as in the logo of the popular Japanese coffee chain “Saint-Marc Café” <サンマルクカフェ> where the last kana carries an acute accent as in the “é”

of the French word “Café,” which is transcribed: the French diacritic is transplanted into the kana syllabary.

Not all combining marks are placed on the same position relatively to the base character, and there are no less than 54 classes of combining characters with respect to the relative position of the diacritic. Such classes are “Above” as in <â>, “Kana_voicing” as in <ポ>, etc.

The rendering of combining character sequences is the responsibility of rendering engines, which combine glyphs in a very precise way, using information stored in the font, namely attachment points placed around glyphs by the font designer.

The second case of character sequencing is the one of *character string*. A character string is a sequence of characters. The order that must be given to characters to obtain graphotactically correct grapheme sequences in the frame of the M-H process is called *logical order*. As U§2.2 puts it:

The order in which Unicode text is stored in the memory representation is called *logical order*. This order roughly corresponds to the order in which text is typed in via the keyboard; it also roughly corresponds to phonetic order.

As hinted by the word “roughly,” there are exceptions to this definition, the most notorious one being the encoding of Thai and Lao scripts: to represent the /ke:/ syllable in Khmer, the logical order agrees with the phonetic order and places the character KHMER LETTER KA <ក< before the character KHMER VOWEL SIGN E <ឺ>, even though the grapheme of the latter is on the left of the grapheme of the former: <ឺក>; to obtain the analogous grapheme sequence in Thai or in Lao, the logical order is to place the character THAI CHARACTER SARA E <๓> (resp. LAO VOWEL SIGN E <ື>) *before* the character THAI CHARACTER KO KAI <ค< (resp. LAO VOWEL KO <ກ<): <ค< (resp. <ກ<) and not *after* the consonant as in Khmer. In other words, logical order agrees with phonetic order in Khmer, but *not* in Thai and Lao, even though these scripts are historically very closely related. The reason is compatibility with preexisting Thai/Lao encodings and typewriter practice.

The greatest advantage of Unicode’s “logical order” is that it solves—at least in computer memory—the problem of mixed left-to-right and right-to-left scripts, such as Latin and Arabic (or Hebrew, or Syriac). In memory, both Latin and Arabic characters are stored in phonetic order. The difficulty arises when such mixed texts have to be displayed. For that, Unicode attaches a default (horizontal) direction to every character: Latin characters have default left-to-right direction (even though Da Vinci wrote the other way around) and Arabic characters have default right-to-left direction. The *Unicode bidirectional algorithm* (Davis, 2019c) provides the order of glyphs for character strings containing characters with different default directions. Because of nested phrases and punctuation marks without default direction, the bidi algorithm sometimes fails to provide the correct result. In that case, the user can insert special characters, such as RIGHT-TO-LEFT EMBEDDING and POP DIRECTIONAL FORMATTING, which will change the algorithm’s output. Here is an example: in the sentence <Did he say “Welcome”?> the question mark is placed outside the quoted <“Welcome”> because it belongs to the noun phrase <Did he say...> and not to the quoted welcome greeting. Translating <“Welcome”> into Hebrew, one gets:

|Did he say ?“ברוך הבא”|,

where the question mark is placed to the left of the quoted phrase, while it should be placed to its right, as it applied to the whole “Did he say ***?” sentence. We avoid this by inserting a LEFT-TO-RIGHT MARK character just before the question mark, resulting in the correct rendering:

|Did he say “ברוך הבא”?|.

This problem would be avoided if there were two distinct exclamation marks in Unicode (a left-to-right one and a right-to-left one), which is

not the case. Only those punctuation marks that have different basic shapes in the two directions have their right-to-left counterparts included in Unicode, e.g., ARABIC QUESTION MARK <؟>, ARABIC COMMA <،> and ARABIC SEMICOLON <؛>.

We can conclude that the formal approach of Sproat (2000) can represent both Unicode combining sequences and character strings, but lacks fine details such as the 54 combining character classes, etc. On the other hand, Unicode logical order comes in handy for people to know in which order they have to type characters, even though it suffers from inconsistencies due to compatibility with legacy encodings. Finally, the Unicode bidirectional algorithm is a good solution for encoding character strings of scripts in different directions, but one needs to take care of ambiguities due to nesting of phrases and to neutral-direction punctuation marks.

5. Ligatures

When a character string is handed over to a rendering engine as input of the [M-H] process, in most cases Sproat's regularity principle applies, so that the derivation of a sequence of adjacent underlying linguistic units is simply the planar catenation of the derivations of individual units. There are nevertheless language-dependent exceptions to this principle, namely *ligatures*.

Ligatures are graphs obtained by merging adjacent graphs. They can be *optional* or *mandatory* (optional in the sense that the adjacent graphs may also, under some conditions, remain unchanged), and their use may or may not be taken into account in linguistic analysis.

- *Mandatory ligatures* are those occurring systematically when two given graphs are adjacent. The most prominent example is the Arabic *lam-alif* |لا| (compare with the hypothetical unligatured *|لا|). The use of the *lam-alif* ligature is a fundamental rule of the Arabic writing system from its very beginnings, as in the following sentence typeset in undotted 6th century CE *Mashq Kufi* (Mousavi Jazayeri, Michelli, and Abulhab, 2017):

لا رأيت ولا سمعت

لا رأيتُ ولا سمعتُ, “I have neither seen, nor heard”). The *lam-alif* ligature is used in *all* Arabic-script languages. It is also noteworthy that it has been taught for centuries in schools as being part of the Arabic alphabet (Dichy, in this volume) and that Arabic typewriters contain a key for it, even though it is not considered as a letter of the Arabic alphabet. Nevertheless, despite its universal presence in the Arabic

script, *lam-alif* remains a ligature and hence there is no *lam-alif* Unicode character²¹.

- *Discretionary ligatures* are those that occur under certain conditions when two given graphs are adjacent. Their use may or not have an incidence on linguistic layers.

We subdivide discretionary ligatures into two classes: *esthetic* ligatures and *linguistically motivated* ligatures (cf. Haralambous 1995):

- *Esthetic ligatures* only contribute to legibility and esthetic quality of the written text. Typical examples are the Latin |fi|, the Arabic |فح| and the Armenian |փւ| (compared with the unligatured |fi|, |مف| and |փւ|). The reasons for using esthetic ligatures are purely visual: to avoid overlapping of bulb and dot in the case of |fi|, to compress text by writing |فح| vertically, to avoid excessive blank space between graphs in the case of |փւ|.

It should be noted that even though esthetic ligatures have no linguistic motivation, their use may be language-dependent. For example, Turkish language does not use ligatures |fi| and |ffi| because the Turkish graphemic system has graphemes <i> and (dotless) <ı>, and the use of the ligatures would cancel their distinctive potential and introduce ambiguity.

- *Linguistically motivated ligatures* have an ambiguous status between stand-alone graphemes and grapheme sequences. Typical examples are the French <œ> and the Dutch <ij>. Both have a grapheme-like behavior when it comes to case, since they are uppercased as stand-alone graphemes: <Œttingen>, <Umegen> (and not *|Oettingen| or *|Ijmegen|). On the other hand, and unlike the Arabic *lam-alif*, they do not appear on typewriters²². They can be qualified as second-class citizens of the graphemic system: they do not appear in prescriptive grammars, are hardly taught in school, and are difficult to access on computer keyboards.

The distinction between esthetic and linguistically involved ligatures can be blurry: for example, in the German language, the f-ligatures are *indirect morphological markers* since they are only used intramorphemically. In German typographic practice, ligatures crossing morpheme boundaries are avoided: |Kaufleute|, |Auffassung|, etc.

21. In fact there is a “presentation form” character ARABIC LIGATURE LAM WITH ALIF ISOLATED FORM, but its usage is highly discouraged: “[Presentation form characters] are included here for compatibility with preexisting standards and legacy implementations that use these forms as characters. Instead of these, letters from the [standard] Arabic block should be used for interchange”. (UŞ9.2)

22. But they were present on the keyboards of localized Monotype/Linotype typesetters in the late 19th and most of the 20th century.

Ligatures are interesting from a theoretical point of view because they challenge the definition of script as a system of distinctive elementary units that allow double articulation. As Nehrlich (2012, p. 30) writes:

Die Ligatur stellt die Hauptmerkmale des Schriftsystems in Frage: Die wohl grundlegendste Anfechtung besteht in der Tatsache, dass das Vorhandensein von Ligaturen das Konzept des Buchstabens problematisiert. Buchstaben sind die Grapheme, aus denen das Alphabet besteht, doch taugen sie ausschließlich als abstrakte Vorstellung. Sobald es um die materielle Realisierung von Schrift geht, verliert der Begriff des Buchstabens an Gültigkeit: Das Vorkommen von Ligaturen falsifiziert die Bestimmung von Buchstaben als das, was innerhalb eines Wortes durch Lücken getrennt ist.²³

Indeed, from a systemic point of view, (esthetic) ligatures are unnecessary since they do not carry any linguistic information, and unnecessary features tend to disappear in an evolving system. But ligatures happen to exist for as long as writing exists and do not seem to face a risk of extinction in the near future. Ligatures make us realize that, just like light has a dual particle/wave nature, graphemes also have a dual nature since they carry both graphical and linguistic information. Similarly to Young's double-slit experiment that has revealed the dual nature of light, ligatures reveal (at least in Western languages) the dual nature of graphemes.

The dual nature of graphemes (and hence also of Unicode characters that represent them in the digital world) is the core difference between M-H and M-M processes, and it comes as no surprise that the Unicode Consortium has very carefully examined the issue of ligatures.

Indeed, Unicode draws a clear line between linguistically motivated ligatures on the one side, and esthetic or mandatory ligatures on the other. The former are first-class citizens of the encoding (for example, <œ> is encoded as LATIN SMALL LIGATURE OE and <ij> as LATIN SMALL LIGATURE IJ). This is not the case of for esthetic and mandatory ligatures²⁴.

Esthetic ligatures are handled by rendering engines, but the user can prevent their use by introducing a special "ligature-breaking" character, called ZERO WIDTH NON-JOINER. This is, for example, what is needed in German language to avoid intermorphemic ligatures.

23. "Ligatures challenge the main characteristics of writing systems: the most fundamental challenge is the fact that the existence of ligatures makes the concept of letter problematic. Letters are graphemes out of which the alphabet is built, but they are valid only as abstract perception. As soon as we deal with material realization of writing, the concept of letter loses its validity: the occurrence of ligatures falsifies the definition of letters as what is separated by gaps inside a word."

24. In fact, many of them do appear in Unicode, but only for reasons of compatibility with legacy encodings, and their use is discouraged.

Contrary to the Latin script, the members of which are usually represented by separate graphs, the Arabic and the Syriac script have two levels of interaction between graphs:

1. on the primary level, a 4-form graph²⁵ is necessarily connected with the graph following it²⁶. Connecting strokes are horizontal and always occur at the baseline, as in |جج|;
2. on a secondary level, discretionary esthetic ligatures occur. In this case graphs are assembled vertically or diagonally, as in |جج| (Haralambous, 1994a).

Since there are two distinct levels of interaction between graphs, one may want to interfere on the first level (separate two graphs that are normally connected) or on the second level (avoid the use of a ligature and return to the standard pair of connected graphs). To allow this two-level interaction, Unicode recommends the use of two distinct characters:

1. the ZERO WIDTH NON-JOINER character (already mentioned above) that acts on the first level and separates graphs by changing their contextual form (the first graph will turn from initial to isolated and from medial to final, the second graph will turn from medial to initial and from final to isolated);
2. the ZERO WIDTH JOINER character that acts on the second level, by preventing any esthetic ligature but keeping the mandatory connection (and therefore not changing the graphs' contextual forms).

As an example, compare the following three:

- standard: جج ;
- with ZERO WIDTH JOINER: ججج ;
- with ZERO WIDTH NON-JOINER: جج·

Contextual form is a graphetic property of Arabic, but in some cases it can contribute to meaning production and can change the status of a grapheme from phonographic to logographic. For example, |ه| (the initial form of grapheme <ه>) is often used as the abbreviation of سنة هجرية "year of the Hegira," and is therefore a logogram. It can also have other meanings: for example, in the French-Arabic dictionary (*Mounged de poche français-arabe* 1991), several abbreviations are written in initial form: |م| for feminine gender (مؤمع), |ج| for plural number (جمع), |ه| for pronouns with nonhuman referents, and the same letter in isolated form |ه| for

25. In the Arabic writing system, graphs |وزرذد| are 2-form graphs (isolated and final form), graph |ه| has only one contextual form and all other graphs are 4-form graphs (isolated, initial, medial and final form).

26. Except for experimental versions of the Arabic script like those described in Haralambous (1998).

pronouns with human referents. Notice that no abbreviation dot is used so that their contextual form is the only indicator of their abbreviative nature.

Transgressing contextual rules for Arabic (or Syriac) graphs is part of the function of the ZERO WIDTH NON-JOINER character: to obtain the (initial) abbreviation |ا| through the M-H process, the ARABIC LETTER HEH character must be followed by the ZERO WIDTH NON-JOINER character. As this operation changes the nature of the grapheme into a logograph with a specific meaning given by the context, the ZERO WIDTH NON-JOINER is necessary also in the frame of the M-M process, even though the machine does not need to visualize Arabic in order to process it.

6. Emojis

Emojis are described in the Unicode Technical Report (Davis, 2019e). The word “emoji” comes from the Japanese 絵文字 (“e” = “picture” and “moji” = “written character”). Emojis are defined in Davis (*ibid.*) as

A colorful pictogram that can be used inline in text. Internally the representation is either (a) an image, (b) an encoded character, or (c) a sequence of encoded characters. (*ibid.*, §1.4)

As many emojis are depicting humans, soon after their introduction questions began to arise about equal depiction of genders, ethnicities, religious minorities, etc. In 2016, the feminine brand *Always* started an advertisement campaign showing young women discussing gender representation in emojis, with slogans such “There aren’t enough emojis to show what girls can do”. To this the then First Lady Michelle Obama replied, on Women’s Day, March 8th, by a tweet:

Hey @Always! We would love to see a girl studying emoji. Education empowers girls around the world. #LetGirlsLearn #LikeAGirl

Following this presidential encouragement to emojis creators and smartphone manufacturers (see Stewart, Maria 2016), the Unicode Consortium faced the problem of sudden emoji multiplication: retaining only masculine white-skin forms was not politically correct, requiring a fixed number of variants for each emoji was unfeasible because of the risk of combinatorial explosion and because whatever the size of the set of variants, it had strong chances of ending up being incomplete in the long run.

The Unicode Consortium adopted a structuralist approach by gradually introducing *dimensions* in the set of emoji variants:

1. the *type of presentation* (typographic B&W or pictorial color);



FIGURE 6. Two ways of rendering emojis: “emoji presentation” on the right, and “text presentation” on the left.

2. *gender*;
3. *color of skin*;
4. *color of hair*;
5. as in Egyptian hieroglyphics, emojis sometimes picture humans or animals *sidewise*. According to cultural conventions (and, in particular, to direction of the dominant writing direction in a given culture), picturing a human or an animal facing/moving to the left or to the right do not have the same connotation, so a fifth dimension has been added: *direction* of sidewise presented human or animal.

To position an emoji in this 5-dimensional space, Unicode provides the mechanism of *emoji sequences*. As with combining sequences, the writer adds, after the “emoji base character” additional Unicode characters corresponding to the intended transformations; rendering engines then, after loading the font, select the appropriate emoji glyph whenever this is possible, or use a fallback mechanism when there is no glyph precisely fulfilling the writer’s demand.

There are five mechanisms allowing to obtain emoji variant glyphs:

1. *presentation sequences*, where a given emoji is followed by VARIATION SELECTOR-15 in order to be presented in B&W typographical style, or followed by VARIATION SELECTOR-16 in order to be presented in colorfull emoji style (see Fig. 6);
2. *modifier sequences*, where an emoji containing some part of human skin is followed by a character that will set the skin color, in five steps from light to dark:

Default image: 🍌; skin color 1: 🍌🏻; 2: 🍌🏼; 3: 🍌🏽; 4: 🍌🏾 and 5: 🍌🏿.

Unicode recommends that the default image (without modifier) should use “a generic, *non-realistic* skin tone” (usually: yellow²⁷);

3. *ZWJ sequences*, where some emojis are connected²⁸ by the ZERO WIDTH JOINER character, as in ligatures. The result of the ZWJ-joining of emojis is implementation dependent: it should result in the rendering of a single emoji incorporating visual elements from all joined emojis.

A recommended use of ZWJ sequences is to have gender appear explicitly in the emoji. For that, there are two mechanisms. Either an emoji depicting a person in a specific role is followed by a ZWJ character and then FEMALE SIGN ♀ or MALE SIGN ♂, or an object is preceded by the emoji MAN or WOMAN and the ZWJ character. Preceding the base emoji by the ADULT 🧑 emoji instead of MAN or WOMAN will produce a gender-neutral appearance. Here is an example: 🧑 is a male worker, this emoji followed by ♂ will remain as is, and followed by ♀ will become 🧑♀.

As can be seen in Fig. 7, in the specific case of the Apple Color Emoji font, the “gender-neutral” ADULT emoji is a morphed intermediate version between MAN and WOMAN, bearing anatomic characteristics and social conventions of both and (according to Western social conventions) having neither a moustache like MAN nor dyed lips like WOMAN.

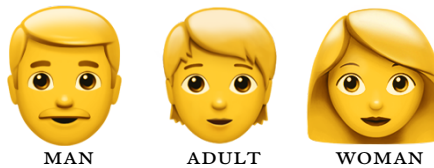


FIGURE 7. A closer look to the “generic” MAN, WOMAN and gender-neutral ADULT emojis, in the Apple Color Emoji font

In a similar manner, one can modify hair color of a face emoji by using ZWJ sequences with emoji components RED-HAIRED, CURLY-HAIRED, WHITE-HAIRED and BALD. Notice that brown/black hair is

27. It can be debated whether *non-realistic yellow* skin color is indeed politically correct, especially when it is combined with blond hair as in the example displayed.

28. While these mechanisms theoretically allow a very wide range of combinations, the Unicode Consortium also publishes Web pages and data files that list the combinations that implementers (font designers and keyboard designers) are expected to support, e.g., <http://www.unicode.org/emoji/charts/full-emoji-list.html#1f46d>. For example, there is an expectation that for a ‘person in lotus position,’ there is also a man and a woman in lotus position, whereas for a ‘person taking bath,’ there is no expectation for gender-specific variants.

default for face emojis, and therefore needs no extra component. Here are the effects of these modifications to the 🧑 emoji: red-haired 🧑🇷🇺, curly-haired 🧑🇨🇷, white-haired 🧑🇵🇪, and bald 🧑🇰🇪.

Finally one can indicate facing direction for emojis displaying humans, by using ZWJ sequences with arrow emojis.

The recommended order of components in an ZWJ sequence is the following: (1) base, (2) emoji modifier or presentation selector, (3) hair component, (4) gender component, (5) direction indicator.

4. A *flag sequence* is a special emoji displaying a black flag 🚩, followed by two ASCII characters representing a country (as in Davis (2019d)). One then obtains the flag of the country, for example 🚩fr should produce 🇫🇷;
5. a *tag sequence* is again a mechanism for obtaining flags, but this time for specific parts of countries, for example Wales as part of the United Kingdom. The approach is different: instead of having the black flag emoji character followed by exactly two ASCII characters, one writes an arbitrary number of “tag characters,”²⁹ and closes the sequence by a specific tag character called CANCEL TAG. The only constraint is that the total length of the sequence (including the black flag and the cancel tag) must be less or equal to 32. So, for example, to obtain the flag of Wales 🇨🇾 one will write 🚩 followed by tag characters gbwls and the CANCEL TAG character;

In linguistic terms, most emojis are pictograms; the exceptions to this rule are mostly cases where symbols are represented, such as 🙅, 🚫, 🚫, etc. In Peirce’s semiotics they are *iconic signs* since they physically resemble their real-world referents. Emojis are included in text, either in an adjunctive or in a substitutive way, and thus contribute to meaning production. Most of them are inherently ambiguous: a smiling face emoji 😊 can mean “I’m happy” or “don’t take it seriously, It’s just a joke” and many other interpretations according to the context. This playful and sometimes poetic ambiguity has certainly contributed to their popularity.

As all Unicode characters, emojis have names such as GRINNING FACE 😊, ROCKET 🚀 or ZOMBIE 🧟. Nevertheless, users are not necessarily aware of names: they choose emojis only according to their shape, and thus attach their own meaning to each emoji. During the act of communication these choices are then confronted to similar choices by other

29. “Tag characters” are ASCII characters transposed to the E0000₁₆ area, in the following sense: if the code point of LATIN SMALL LETTER A is hexadecimal 61₁₆ (that is decimal 97), then the code point of the corresponding TAG LATIN SMALL LETTER A character is E0000₁₆ + 61₁₆ = E0061₁₆ (decimal 917,601). In this text we will represent tag characters by underlined typewriter glyphs to prevent confusion with ordinary ASCII characters.

people, and this process results in a series of constantly evolving consensuses. In addition to that, on every new smartphone system release, a few hundred new emojis are added, enlarging the semantic spectrum available to users. C. Servais and V. Servais (2009) claim that “misunderstanding is the basic pattern of communication,” this is even more true when we consider communication by emojis.

To cut short this situation of ambient ambiguity and to solve once and for all the problem of emoji proliferation, the President of the Unicode Consortium, Mark Davis, submitted a groundbreaking proposal for indefinitely extending the number of emojis while precisely pinpointing their semantics.

The QID Emoji Proposal

The proposal (Davis, 2019a) was submitted to the Unicode Technical Committee on May 2nd, 2019, and at the time of writing of this paper it is not known whether it will be accepted.

Before describing the proposal, let us introduce the *Wikidata Project*. Wikidata is a collaborative knowledge base. It was launched by the Wikimedia Foundation in October 2012.

Wikidata has a graph³⁰ structure with *items*, *literals* and *media files* as vertices, and *statements* as edges. Items can be topics, concepts or objects. Statements connect items between them, items with literals (character strings or numbers), or items with media files. Each statement is an instance of a *property*. Each item has an identifier: the letter “Q” followed by a number; each property has an identifier: the letter “P” followed by a number. Statements may have *qualifiers* which are additional pieces of information. As of today (October 16th, 2019) Wikidata contains 63,573,864 data items and 6,762 properties.

As an example, the city of Brest (located in Brittany, France) is represented by item Q12193. Here are some of its statements:

Property	Value
P31 (instance of)	Q484170 (commune de France)
P31 (instance of)	Q1549591 (big city)
P17 (country)	Q142 (France)
P1313 (office held by head of government)	Q62266917 (Mayor of Brest)
P6 (head of government)	Q3084338 (François Cuillandre)


30. In this section the term “graph” refers to the mathematical structure (a set of binary relations) and *not to* the elementary material unit of writing, as in the rest of the paper.

By following these links we find out that the item “Brest” represents a big city in France, governed by a “Mayor Of Brest,” and this position is occupied by the referent of item Q3084338, called “François Cuillandre”. These are only 5 among the 136 statements provided for the item “Brest” in Wikidata.

Wikidata follows an intensional approach to information: items of the real world are entirely represented by their properties. These properties link items with other items, building a graph of relations between them. A human can retrieve information by following the relations of this graph and an inference engine can reply to queries formulated by humans.

Let us now describe the QID Emoji Proposal: Mark Davis proposes the establishment of a *one-to-one correspondence between emojis and Wikidata items*³¹. On a technical level, every emoji would be identified by a new kind of tag sequence, starting with a special generic emoji EMOJI TAG BASE, followed by a Wikidata QID identifier written in tag characters, and finally a CANCEL TAG character. For example, an emoji for the town of Brest would be obtained by the tag sequence [EMOJI TAG BASE]Q12193[CANCEL TAG].

The consequences of this initiative, if it is adopted, can be important. On the technical level, the size of the set of Unicode-encoded emojis will go from a few thousands to more than 60 millions. Smartphone providers will need to invent new ways of sharing fonts on the Web to provide emoji glyphs to any user requesting them—and for emojis not yet drawn, fallback glyph selection algorithms have to be applied.

But the most important consequence will be on the level of human communication: the new kind of emojis will be *significantly less ambiguous than written text*. For example, the textual sentence “I live in Brest” is ambiguous since there are at least 9 towns or villages with that name in the world (in Belarus, Bulgaria, Croatia, Czech Republic, France, Germany, Poland, Serbia and Slovenia), but the sentence “I live in [Q12193 emoji]” (potentially displayed as “I live in ”) is unambiguous³².

Furthermore we will witness a progressive shift from process M-H to process M-M: in process M-M the machine identifies concepts in linguistic data and replaces them with, for example, Wikidata identifiers. Using QID emojis in the M-H process, one will get a visual result similar to the existing one, but the Unicode data used to obtain it will already

31. With the possible exception of existing emojis, for which we don’t know whether they will be assigned to QIDs.

32. It is the code “Q12193” which is unambiguous, not the image of the emoji, where we see a tower that probably only an inhabitant or native of Brest will recognize as being the Tour Tanguy.

contain the necessary information for the NLP algorithm to unambiguously identify the meaning of the emoji that is part of the text.

For the moment, Wikidata items are only nouns, but one may very well imagine an extension to verbs (similar to WordNet, which has sections for nouns, verbs, adjectives and adverbs). This would allow the replacement of the verb “I live” by an “extended-QID” emoji, so that all lexical morphemes of the sentence “I live in Brest” are replaced by references to Wikidata. This process, known as *semantic annotation*, is very common in Artificial Intelligence.

Considering QID emojis with a large amount of optimism, one could say that, thanks to them, semantic annotation will become part of everyday human communication, and this may very well result in being a major turning point in human communication. But in fact, our optimism is limited since QID emojis could also create a range of problems and misunderstandings:

1. QIDs imply well-curated semantics, but emojis may quickly be repurposed. As an example, the peach emoji was already overloaded with meaning beyond that of the literal fruit. But in fall 2019, in a matter of weeks if not days, it acquired an additional meaning of “impeach,” based on the sudden prominence of the political topic in the USA and the phonetic similarity. Any hope of keeping the meaning of any emoji in any way limited to that of the underlying QID seems totally hopeless from the very start. There is no “emoji police,” and writers use emoji based on appearance, imagination, and consensus, rather than based on name or formal definition.
2. As a consequence of the previous point, it would be impossible for NLP software to put too much confidence in any kind of QID being used in an emoji. In many cases, somewhat paradoxically, deriving semantics from words (such as “peach”) might be considerably easier than deriving semantics from an emoji with a QID.
3. Although this may seem implied by the use of a QID grounded in ontology, there is no guarantee that a particular such emoji would be recognized as a depiction of the intended signifier. As an example, even the most prominent building or monument standing for (French) Brest may not be known to a wide range of people, even if these people have no problem to quickly identify Brest as a city in France.
4. While emojis are not specified to a single design, for many of them, the design is informed by the proposals made during the approval process and by the files depicting the newly accepted emojis. Major changes in interpretation, such as when the design of the pistol emoji was changed from a handgun to a water pistol (ABC News, 2018), happen only rarely. With QID emojis, if two people independently create emojis for the same QID, there is no guarantee that there is any kind of image similarity between the two emojis.

5. QID emojis may give the impression that literally everybody can start to use an emoji for any kind of concept. But experience with encoding existing minor scripts has shown that it is very difficult to make sure that the necessary fonts are widely available, even for well-defined language communities. And “install this font to read this Webpage” is a more realistic request than “install this font to view this emoji”. So realistically, QID emojis can only be introduced by major vendors, i.e., the groups that currently publish emoji fonts.
6. Emoji demand exists not only for well-defined ontological concepts, but also for combinations of concepts (e.g., cat with smiling face and tears). Such emojis cannot be created using QIDs, unless Wikidata gets diluted with such combinations.
7. Because each tag character needs four bytes for encoding, whereas ASCII characters need only one byte, it can easily be more efficient to use markup to add ontologically grounded meaning to text (including emojis) than to combine the meaning layer and the appearance in a single code. Using markup to add meaning also leads to a clear separation of concerns and a general solution (because it works for all kinds of text, not only a subset of emojis).

Conclusion

As an encoding, Unicode is a pervasive technology which probably will continue to exist as long as humanity will use text, be it in material or in disembodied form. But Unicode also provides a framework for the descriptive analysis of writing systems, which deserves to be scrutinized from a linguistic point of view, and this is what we attempted. We hope that this will be the starting point for research that will bring the community of Unicode aficionados and the community of (grapho)linguists closer together, and will result in a better understanding of the rationale of this wonderful human achievement.

References

- ABC News (2018). “Gun Emoji Replaced with Toy Water Pistol across All Major Platforms”. <https://perma.cc/V8DQ-ANKN>.
- André, Jacques and B. Borghi (1990). “Dynamic Fonts”. In: *PostScript Language Journal* 2.3, pp. 4–6.
- Anis, Jacques (1988). *L'écriture, théorie et descriptions*. Bruxelles: De Boeck.
- Bao, Jie et al. (2012). “OWL 2 Web Ontology Language Document Overview”. <https://www.w3.org/TR/owl2-overview/>.

- Bayar, Abdelouahad and Khalid Sami (2010). "Towards a Dynamic Font Respecting the Arabic Calligraphy". In: *Handbook of Research on E-services in the Public Sector: E-government Strategies and Advancements*. Ed. by Abid Thyab Al Ajeeli and Yousif A. Latif Al-Bastaki. Hershey PA: IGI Global, pp. 359–379.
- Bellamy-Royds, Amelia et al. (2018). "Scalable Vector Graphics (SVG) 2". <https://www.w3.org/TR/SVG2/>.
- Berglund, Anders (2006). "Extensible Stylesheet Language (XSL) Version 1.1". <https://www.w3.org/TR/xs111/>.
- Bray, Tim et al. (2008). "Extensible Markup Language (XML) 1.0". <https://www.w3.org/TR/xml/>.
- Brown, Keith and Jim Miller (2013). *The Cambridge Dictionary of Linguistics*. Cambridge: Cambridge University Press.
- Carlisle, David, Patrick Ion, and Robert Miner (2014). "Mathematical Markup Language (MathML) Version 3.0". <https://www.w3.org/TR/MathML3/>.
- Chomsky, Noam and M. Halle (1968). *The Sound Pattern of English*. New York: Harper & Row.
- Daniels, Peter and William Bright (1996). *The World's Writing Systems*. 2nd ed. Oxford: Oxford University Press.
- Davis, Mark (2019a). "QID Emoji Proposal". <http://www.unicode.org/L2/L2019/19082r-qid-emoji.pdf>.
- (2019b). "Unicode Standard Annex 29. Unicode Text Segmentation". <https://www.unicode.org/reports/tr29/>.
- (2019c). "Unicode Standard Annex 9. Unicode Bidirectional Algorithm". <https://www.unicode.org/reports/tr9/>.
- (2019d). "Unicode Technical Standard 35. Unicode Locale Data Markup Language". <https://www.unicode.org/reports/tr35/>.
- (2019e). "Unicode Technical Standard 51. Unicode Emoji". <http://www.unicode.org/reports/tr51/>.
- Dichy, Joseph (in this volume). "On the Writing System of Arabic. The Semiographic Principle as Reflected in Nashī Letter Shapes".
- Dukes, Kais, Eric Atwell, and Nizar Habash (2013). "Supervised Collaboration for Syntactic Annotation of Quranic Arabic". In: *Language Resources and Evaluation* 47.1, pp. 33–62.
- Dürscheid, Christa (2016). *Einführung in die Schriftlinguistik*. 5th ed. Göttingen: Vandenhoeck & Ruprecht.
- (2018). "Bild, Schrift, Unicode". In: *Sprache – Mensch – Maschine. Beiträge zu Sprache und Sprachwissenschaft, Computerlinguistik und Informationstechnologie für Jürgen Rolsboven aus Anlass seines sechsundsechzigsten Geburtstages*. Ed. by Guido Mensching et al. Köln: Kölner UniversitätsPublikationsServer, pp. 269–285.
- Dürst, Martin and Asmus Freytag (2000). "Unicode in XML and Other Markup Languages". <https://www.w3.org/TR/2000/NOTE-unicode-xml-20001215/>.

- Dürst, Martin and M. Suignard (2005). “Internationalized Resource Identifiers (IRIs)”. Request for Comments 3987.
- Gompel, Maarten van and Martin Reynaert (2013). “FoLiA: A Practical XML Format for Linguistic Annotation—a Descriptive and Comparative Study”. In: *Computational Linguistics in the Netherlands Journal* 3, pp. 63–81.
- Grzybek, Peter and Milan Rusko (2009). “Letter, Grapheme and (Allo-)Phone Frequencies: The Case of Slovak”. In: *Glottology* 2, pp. 30–48.
- Günther, Hartmut (1988). *Schriftliche Sprache: Strukturen geschriebener Wörter und ihre Verarbeitung*. Tübingen: Niemeyer.
- Haralambous, Yannis (1991). “Typesetting Old German: Fraktur, Schwabacher, Gotisch and Initials”. In: *TUGboat* 12.1, pp. 129–138.
- (1994a). “The Traditional Arabic Typecase Extended to the Unicode Set of Glyphs”. In: *Electronic Publishing—Origination, Dissemination, and Design* 8.2/3, pp. 125–138.
- (1994b). “Typesetting Khmer”. In: *Electronic Publishing—Origination, Dissemination, and Design* 7.4, pp. 197–215.
- (1995). “Tour du monde des ligatures”. In: *Cahiers GUTENBERG* 22, pp. 69–70.
- (1998). “Simplification of the Arabic Script: Two Different Approaches and Their Implementations”. In: *Springer Lecture Notes in Computer Science*. Vol. 1375: *Electronic Publishing, Artistic Imaging, and Digital Typography*, pp. 138–156.
- (2007). *Fonts & Encodings. From Advanced Typography to Unicode and Everything in Between*. Sebastopol, CA: O’Reilly.
- (2013). “New Perspectives in Sinographic Language Processing Through the Use of Character Structure”. In: *Springer Lecture Notes in Computer Science*. Vol. 7816: *CICLing 2013: 14th International Conference on Intelligent Text Processing and Computational Linguistics, Samos*, pp. 201–217.
- Hayes, Patrick J. and Peter F. Patel-Schneider (2014). “RDF 1.1 Semantics”. <https://www.w3.org/TR/rdf11-mt/>.
- Hellwig, Oliver (2010–2019). “DCS—The Digital Corpus of Sanskrit”. <http://www.sanskrit-linguistics.org/dcs/>.
- Heninger, Andy (2019). “Unicode Standard Annex 14. Unicode Line Breaking Algorithm”. <https://www.unicode.org/reports/tr14/>.
- Kaplan, Ronald M. and Martin Kay (1994). “Regular Models of Phonological Rule Systems”. In: *Computational Linguistics* 29, pp. 331–378.
- Kazuo, Inamori [稲盛和夫] (2019). 心 [The Mind]. Tokyo: サンマーク出版 [Sunmark Publishing].
- Lotman, Jurij (1977). *Michigan Slavic Contributions*. Vol. 7: *The Structure of the Artistic Text*. Ann Arbor: The University of Michigan.
- Mackenzie, Charles E. (1980). *Coded Character Sets, History and Development*. Reading, MA: Addison-Wesley.

- Marneffe, Marie-Catherine de et al. (2013). “More Constructions, More Genres: Extending Stanford Dependencies”. In: *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*. Prague: Matfyzpress, pp. 187–196.
- Martinet, André (1970). “La double articulation du langage”. In: *La linguistique synchronique*. Paris: PUF, pp. 7–41.
- Meletis, Dimitrios (2015). *Graphetik. Form und Materialität von Schrift*. Glückstadt: Verlag Werner Hülsbusch.
- (2019). “Naturalness in scripts and writing systems: Outlining a Natural Grapholinguistics”. PhD thesis. University of Graz.
- Mounged de poche français-arabe* (1991). Beyrouth: Dar el-Machreq.
- Mousavi Jazayeri, S.M.V., Perette E. Michelli, and Sadd D. Abulhab (2017). *A Handbook of Early Arabic Kufic Script*. New York: Blautopf Publishing.
- Myers, James (2019). *The Grammar of Chinese Characters. Productive Knowledge of Formal Patterns in an Orthographic System*. London, New York: Routledge.
- Nehrlich, Thomas (2012). “Phänomenologie der Ligatur. Theorie und Praxis eines Schriftelements zwischen Letter und Lücke”. In: *Von Lettern und Lücken: zur Ordnung der Schrift im Bleisatz*. Ed. by Mareike Giertier and Rea Köppel. Paderborn: Wilhelm Fink, pp. 13–38.
- Pemperton, Steven et al. (2018). “XHTML 1.0 The Extensible HyperText Markup Language”. <https://www.w3.org/TR/xhtml1>.
- Rezec, Oliver (2009). “Zur Struktur des deutschen Schriftsystems. Warum das Graphem nicht drei Funktionen gleichzeitig haben kann, warum ein <a> kein <a> ist und andere Konstruktionsfehler des etablierten Beschreibungsmodells. Ein Verbesserungsvorschlag”. PhD thesis. Ludwig-Maximilians-Universität Munich.
- Sawicki, Marcin et al. (2001). “Ruby Annotation”. <https://www.w3.org/TR/2001/REC-ruby-20010531/>.
- Schopp, Jürgen F. (2008). “In Gutenbergs Fußstapfen: Translatio typographica. Zum Verhältnis von Typografie und Translation”. In: *Meta* 53, pp. 167–183.
- Servais, Christine and Véronique Servais (2009). “Le malentendu comme structure de la communication”. In: *Questions de communication* 15, pp. 21–49.
- Sproat, Richard (2000). *A Computational Theory of Writing Systems*. Cambridge: Cambridge University Press.
- Stewart, Maria (2016). “Michelle Obama Just Suggested a New Emoji to Empower Girls”. https://www.huffpost.com/entry/michelle-obama-always-emoji_n_56df3feee4b03a40567a78c3?guccounter=1.
- Stöckl, Harmut (2004). “Typographie: Gewand und Körper des Textes – Linguistische Überlegungen zu typographischer Gestaltung”. In: *Zeitschrift für angewandte Linguistik* 41, pp. 5–48.

- The Unicode Standard. Version 12.0—Core Specification* (2019). Mountain View, CA: The Unicode Consortium.
- Wmffre, Iwan (2008). *Contemporary Studies in Descriptive Linguistics*. Vol. 23: *Breton Orthographies and Dialects: The Twentieth-Century Orthography War in Brittany*. Bern: Peter Lang.


Emojis: A Grapholinguistic Approach


Christa Dürscheid & Dimitrios Meletis

Abstract. The present article stands at the interface of CMC research and grapholinguistics. After outlining which features are typical of the writing of private text messages, the focus of the first part of the paper (Sections 2 and 3) lies on the use of emojis. Notably, emoji use is not—as is commonly done—analyzed under a pragmatic perspective, but grapholinguistically, at the graphetic and graphematic levels: emojis are conceptualized as visual shapes that may assume graphematic functions within a given writing system. In the second part (Section 4), it is underlined that all variants of written digital communication (such as the use of emojis, but also all other characters) are made possible only due to the Unicode Consortium’s decisions; this, finally, is argued to have far-reaching consequences for the future of writing.

1. Preliminary Remarks

In this paper, the use of emojis will be considered within a framework known in the German-language research area as “Schriftlinguistik” (*grapholinguistics*). As will be demonstrated, this term is not equivalent to the terms *graphemics* or *graphematics*. In a much broader sense, grapholinguistics entails different aspects of writing (among them research on scripts and writing systems, the history of writing, orthography, graphematics, the acquisition of reading and writing, text design and text-image-relations, and differences between the written and spoken modalities of language) (cf. Dürscheid 2016).¹ This paper’s main

Christa Dürscheid  0000-0001-9141-7562
Department of German Studies, University of Zurich
Schönberggasse 9, 8001 Zürich, Switzerland
duerscheid@ds.uzh.ch

Dimitrios Meletis  0000-0002-8889-6459
Department of Linguistics, University of Graz
Merangasse 70/III, 8010 Graz, Austria
dimitrios.meletis@outlook.com

1. To date, this textbook is only available in German (in its 5th edition).

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1.
Fluxus Editions, Brest, 2019, p. 167–183. <https://doi.org/10.36824/2018-graf-duer>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

focus will be on a certain phenomenon within this vast field of topics—the fact that texts are increasingly being enriched by images. These include emojis,² ASCII signs, stickers, GIFs, photos, and videos, i.e., different kinds of visual elements that Herring and Dainas (2017) subsume under the umbrella term *graphicons*. Among these graphicons, emojis constitute their own inventory of visual units. Not only is their number growing annually (at this point, there exist about 3,000, see <https://unicode.org/emoji/charts/emoji-counts.html> <31.08.2019>), but their use in everyday writing, for instance in WhatsApp messaging, is also on the rise. Unlike, for example, photos or videos, emojis function as an integrated part of text messages. They are situated on the same line as the other characters and often substitute them (cf. *I'll come by car* > *I'll come by 🚗*). These features give rise to the question whether emojis may become the basis of a new way of writing (or even a new language), a question never asked with regard to the other types of graphicons. This question is also motivated by the unique technical status of emojis: among graphicons, they are the only visual elements that are included in the Unicode Standard. Notably, the inclusion of each new emoji requires a well-elaborated proposal to the Unicode Consortium. However, once such a proposal is approved, the emoji in question can be inserted into texts like any other character (see Section 4).

The theoretical framework on which this paper is based will be discussed in the next section: we will present relevant research on computer-mediated communication (CMC) on the one hand and on grapholinguistics on the other. After that, a short overview of emoji research will be given. Here, the focus will shift towards the question of how emojis may be analyzed from a grapholinguistic point of view (Section 3). In this context, data from a Swiss project empirically investigating the use of emojis will provide insight into the various functions they fulfill in WhatsApp messages (cf. Ueberwasser and Stark 2017). While these functions can be explained from a pragmatic perspective (cf. Danesi 2016; Pappert 2017; Beißwenger and Pappert 2019; Dainas and Herring in press), the present paper will instead focus on the functions emojis fulfill at the graphematic level (cf. Dürscheid and Frick 2016, Dürscheid and Siever 2017). Section 4 will then address the question of which role the Unicode Consortium plays with respect to the use of emojis. How far-reaching are the consortium's decisions and what are the consequences of the (non-)inclusion of a graphic sign in the Unicode

2. As for the plural of *emoji*, the Oxford English Dictionary states that both variants, *emoji* and *emojis*, are allowed (see <https://www.oed.com/>). Interestingly, in 2016, Emojipedia, a famous website covering the use of emojis, revealed that, based on empirical data, the use of plural-*s* is increasingly popular (see <https://blog.emojipedia.org/emojis-on-the-rise-as-plural/> <30.09.2019>).

character set? A short reflection on the future of emoji use and related open questions will conclude the paper (Section 5).

2. Theoretical Background

Significant research on CMC is closely linked to the name of Susan Herring, Professor of Information Science and Linguistics at the Indiana University Bloomington, where she also founded and still directs the Center for Computer-Mediated Communication. Of her many works on the topic, one that is particularly worth mentioning is “Pragmatics of Computer-Mediated Communication,” a handbook she co-edited with two colleagues (cf. Herring, Stein, and Virtanen 2013). In his chapter, Markus Bieswanger compiles the most relevant features of writing in CMC and discusses them at both the grapholinguistic level and the stylistic level (cf. Bieswanger 2013). Bieswanger lists a bundle of typical writing features for CMC such as acronyms (*OMG*), letter and number homophones (*4you*), nonstandard spellings, and punctuation (*really???*). As far as the stylistic level is concerned, he describes, among other features, the accumulation of syntactic reductions and the use of colloquial expressions or dialectal elements. It is noteworthy that these features are used predominantly in private, informal everyday communication (e.g., messages in a WhatsApp family chat). Obviously, this means that not all types of texts on the internet exhibit these features. For example, to date, they hardly ever occur in texts directed at a large, anonymous readership (e.g. on university and company websites) or texts produced in the context of more formal one-to-one communication (e.g., business emails).³

While the features listed above are discussed in detail in both German and English research on CMC, a different approach is found predominantly in the German research tradition: Here, a terminological distinction is made between *medium*, *form of communication*, and *text genre* (cf. Dürscheid 2005). A letter of application, for instance, can be considered a special type of text (*text genre*) that may be sent as an email (*form of communication*) via computer or mobile phone (*medium*). However, the boundaries between these devices are increasingly blurred, as nowadays, mobile phones function almost identically to computers and can be used to write a range of significantly differing types of texts such as letters of application or Facebook postings (for example about one’s last holiday trip); these, ultimately, constitute texts from entirely different *text genres*. The term *form of communication* is used to describe the various

3. This applies to the first contact with customers. If emails are exchanged back and forth quickly, formalities may be abandoned to some extent. This is to say that the more dialogical a text becomes, the sooner the above-mentioned features occur.

communicative practices which are possible within these media. These include an email exchange, a telephone call, a text chat, or any other kind of interaction at the oral or the written level (cf. Jucker et al. 2018). *Text genre*, finally, refers to different communicative purposes that motivate these interactions and enable different types of written texts (or different types of oral conversations, respectively). Some examples for such text genres are (at the written level) business letters, love letters, letters of application, or holiday greetings. Among the given examples, it is predominantly the area of CMC research meeting the following criteria that is treated in this paper: texts which are mediated by smartphones and are part of an interpersonal exchange carried out in a private, informal setting. Consequently, *text genres* such as business letters are not taken into consideration here, and neither are more formal communications on LinkedIn or other social networks.

As mentioned above, we will concentrate on the analysis of the graphematic functions of emojis, which means that the following considerations are situated at the interface between CMC and grapholinguistics. The term *grapholinguistics* is used here instead of other alternatives such as *graphonomy* or *grammatology* which are meant to designate research on writing systems (cf. the numerous works of Peter T. Daniels and Florian Coulmas, for instance). One reason for insisting on grapholinguistics is that we need an expression that refers not exclusively to one research domain of written language but to *all* writing-related aspects (cf. Dürscheid 2016). Furthermore, the use of *grapholinguistics* is of programmatic character, highlighting that writing is by no means a secondary system subordinate to spoken language but instead a fully functional form of language in and of itself and must be examined in its own terms (cf. also Meletis 2019). Worth mentioning in this respect is a dictionary of “Schriftlinguistik” edited by Martin Neef, Said Sahel, and Rüdiger Weingarten. It is part of a series of online (and, later, printed) dictionaries covering various linguistic subfields (e.g., phonetics and phonology, word formation). While this project started out in German, the long-term plan is to also publish the dictionaries in English. The fact that grapholinguistics is a field included in this compilation of dictionaries indicates that its relevance in German linguistics has risen. This is also underlined by the fact that more and more research is being embedded in this framework (cf. Neef 2015; Meletis 2018; Dürscheid 2018).

Interestingly, since 2009, there has even been an entry on grapholinguistics in the German Wikipedia (see <https://de.wikipedia.org/wiki/Schriftlinguistik>, <30.09.2019>). The English Wikipedia, on the other hand, only includes an entry on *graphemics* but not *grapholinguistics*. It states that “graphemics or graphematics is the linguistic study of writing systems and their basic components, i.e., graphemes” (

wikipedia.org/wiki/Graphemics, <30.09.2019>).⁴ This gives readers the impression that *graphematics* encompasses all aspects of the study of writing systems which is, however, inaccurate: writing systems research deals with many more topics than graphematics—and grapholinguistics is still broader (cf. Meletis 2019, Chapter 2). Meletis distinguishes between *graphetics*, the study of the visual resources used in writing, and *graphematics*, the study of the relation between visual units (so-called “basic shapes”) and corresponding linguistic units (such as phonemes, syllables, morphemes). While graphetics treats all aspects of the materiality of writing (as, for example, the choice of typeface or the effect its appearance has on its processing by humans), its “main object of study is *scripts*, defined as inventories of discrete visuo-graphic basic shapes such as the Roman script, the Chinese script, and the Japanese inventories hiragana and katakana” (Meletis, 2018, p. 62). These scripts and the basic shapes they consist of—in the case of Roman script often referred to as ‘letters,’ in Chinese script as ‘characters,’ but cf. Meletis (in press)—are studied for their materiality alone, i.e., dissociated from any linguistic function they might assume in a given context. They are not bound to a given language and its respective writing system, which becomes obvious when considering that many of them—such as the Roman script and the Cyrillic script—are commonly used for more than one writing system (e.g., English, German, Dutch, Italian, and many more for Roman, and Russian, Ukrainian, Serbian, etc. for Cyrillic).

Following this view, a writing system, as the main object of study of graphematics, is the combination of a script and a language (cf. Weingarten 2011). Thus, for instance, the German writing system employs Roman script for the German language, the English writing system Roman for English, the Ukrainian writing system Cyrillic for Ukrainian. The inventory of punctuation signs could also be seen as a script, as could the inventory of digits. Both of these inventories are employed across an even wider range of writing systems than scripts such as Roman or Cyrillic; consider, for example, the comma which appears in very similar functions in many typologically diverse writing systems. Similarly, in our grapholinguistic approach, emojis constitute their own inventory of basic shapes and are used as communicative and sometimes genuinely graphematic resources whose functions are not specific to a given language or writing system, although this would have to be tested in a comparative typological study. The different facets of emoji use will be explored in the following section.

4. Note that on the website of the conference at which a part of this paper was presented, both terms are also used as synonyms: while the conference title was “/gɾafematik/,” its subtitle was “Graphemics in the 21st century” (see <http://conferences.telecom-bretagne.eu> <25.09.2019>).

3. Emojis and Their Use

In the last years, we witnessed a rise in works on emojis specifically from linguistic and semiotic perspectives. The following selection of titles is thus only supposed to give a first idea of the current state of research in this field: Marcel Danesi's book "The Semiotics of Emoji" (2016) distinguishes between emoji semantics, emoji grammar, and emoji pragmatics, demonstrating the use of emojis from these different perspectives. Susan Herring and Ashley Dainas examine different types of graphicons (among them emojis) sampled from public Facebook groups and analyze their frequency as well as their pragmatic functions (cf. Herring and Dainas 2017). According to their findings, emojis may serve, for instance, to express feelings or to clarify the communicative intention of an utterance (as a kind of "tone modification"). Another article (cf. Dainas and Herring in press) presents an emoji survey "administered online in early 2018 to determine how social media users interpret the pragmatic functions of popular emoji types". The abstract from which this quote is taken concludes with the assertion of "the importance of analyzing emoji meaning from the perspective of pragmatics". A concise monograph that strongly emphasizes this aspect has just been published in German and is titled "Handeln mit Emojis" (Beißwenger and Pappert, 2019).⁵ In it, the authors distinguish two main strategies of emoji use: *making readable* ("Lesbarmachen") and *making visible* ("Sichtbarmachen"). *Making readable* refers to using emojis in order to provide readers with information on how to interpret an utterance, while the goal of *making visible* is visually framing an utterance (cf. *ibid.*, pp. 71–73).

While all of the above-mentioned works are grounded in semiotic or pragmatic approaches, Dürscheid and Siever (2017) focus on the grapholinguistic functions that emojis fulfill. Graphetically, they can be used as visual units to separate sentences from each other (instead of a period or a comma) or to indicate the end of the message, and graphematically, they can be functionalized in order to substitute a single grapheme or a sequence of graphemes. Note that this structural analysis of emojis does not compete with the determination of their communicative functions but instead complements the pragmatic approach with a different perspective. This is illustrated in Fig. 1, a text message sent along with a photo.

As is evident from this example, emojis are positioned on the same line as characters and are approximately equal in size. The photo, on the contrary, is presented separately. Although it is semantically connected with the text, it is not positioned within the text, but on top of it. The text itself consists of a short sentence followed by five sun emo-

5. The English translation is (the authors' own suggestion): "How to do things with emojis".



FIGURE 1. Text message with emojis

jis which likely imply that the sun is shining wherever the message was composed. It is also possible, however, that the writer used the sun emoji only in order to render the message a little more cheerful and colorful (cf. Dürscheid and Frick 2016), i.e., with no intention at all of making a statement about the current weather situation. Of course, it is also possible that the writer wanted to combine these two functions. Irrespective of these considerations, it must be noted that the five emojis in 1 do by no means stand for the word *sun*, which is to say that they are not used logographically. If this were the case, the text would have to be read as *The beach says hi sun sun sun sun sun*, and it is highly unlikely that this was the writer's intention. Thus, in this example, the sun emoji is used merely as a graphetic resource that does not assume any linguistic function, i.e., it does not refer to any specific linguistic unit. Irrespective of this, it does of course have a context-sensitive communicative function.

In the following, however, we will show that emojis, similar to other basic shapes, can be graphematically functionalized in order to refer to different linguistic levels: In the word *month*, for example, the sun emoji may replace the <o>, i.e., be used as an allograph of the letter <o>. For the word *frontdoor*, an emoji representing a door can be used to replace <door> (*front* 🚪). The emoji in the sentence *Shall we build a 🧑‍🌾 today?* in which it substitutes the word *snowman* functions similarly. If the writer were to also omit the article in this sentence, the emoji would even substitute an entire noun phrase (i.e., *a snowman* or *the snowman* or *our snowman*). As this example shows, interpreting sentences in which an emoji substitutes a noun phrase might produce a number of different readings. Technically, in the last two examples, emojis function as ideograms.⁶

6. In this vein, emojis are similar not only to digits but also to other special characters such as <%> or <&>.

The concept of ideography—at least the question of whether it constitutes writing—has been under a lot of scrutiny (cf. Unger 1990), and today, it is common consensus that ideograms are not considered writing since in its narrow definition, writing is interpreted only as the graphic representation of *specific* linguistic units (cf. Daniels 2018, p. 157). Following this, only *glottography*, i.e., ‘language writing,’ is considered writing, contrary to what is known as *semasiography*, referring to visual units that represent concepts or ideas (cf. *ibid.*, p. 126). While glottography can be read, i.e., decoded directly, semasiography can only be interpreted but never read since no specific linguistic units are associated with the visual shapes.

Returning to our example from above, it is not clear how the snowman emoji would be spelled out. This means that strictly speaking, emojis are ideograms, visual resources used with a communicative function and a meaning, but they are not writing *proper*. This, however, would restrain us from analyzing them in a grapholinguistic approach given that grapholinguistics is only invested in the study of writing. The solution is that while at the formal level, emojis are special characters, graphematically they may be ideograms, and in some uses even logograms: when decoding a given written utterance in which an emoji substitutes a phrase, a word, or a morpheme, the reader commonly decides for one *specific* phrase, word, or morpheme, respectively, in order to read the utterance. This is, for instance, the case in the example given in the preliminary remarks above, *I’ll come by car* > *I’ll come by* 🚗. Here, the emoji is associated with a specific linguistic unit—in our terms, it is used graphematically. Note that such an association is often fluid and not only different individuals might associate emojis with different linguistic units, but also the same individual might read the same emoji in different ways depending on various contextual factors.

Authentic examples of the different uses of emojis both at the segmental and suprasegmental graphematic levels are presented in Dürscheid and Siever (2017).⁷ This paper also provides data on how often emojis are used in WhatsApp chats and which of them are the most popular. This analysis is based on a research project on WhatsApp communication in Switzerland (see <http://www.whatsup-switzerland.ch> <30.09.2019>). The data were collected in 2014, and the text corpus consists of around 750,000 messages for which sociodemographic information is also available (age, gender, mother tongue, etc.). In Fig 2, one example chosen from this corpus will be presented.

The text in the example approximately translates to “I already went ___ today and the weather is nice for once,” where the underscore indi-

7. A short English version of this paper, titled “Beyond the alphabet—communication with emojis,” can be found on www.academia.edu.

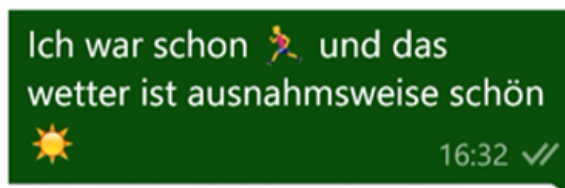


FIGURE 2. Graphematic use of the ‘Person Running’ emoji

cates the first emoji⁸ which replaces a graphematic word, i.e., a sequence of graphemes—the question, now, is which one, as this emoji (its name on Emojipedia is given as “Person Running,” its Unicode code point is U+1F3C3) can represent a variety of verbal expressions: *running*, *on the run*, *walking*, or *jogging*, to name a few. Thus, as illustrated in Fig. 3, the emoji as a visual unit is the *signifier* of different morphemes which are themselves, in the sense of Saussure, bilateral signs. This renders the emoji graphematically ambiguous, as the specific linguistic unit it refers to is not fixed but variable and determined by the context or the reader’s interpretation of a given text in which it is used. Note that the global concept the emoji represents—in the case of Fig. 3 the common concept underlying the words *run*, *walk*, *jog*, and others—is relatively constant and allows the emoji to be interpreted irrespective of the given context. When used to substitute morphemes, words, phrases, etc. within written sentences, emojis become graphematic units as they are treated by readers as sequences which are *read* instead of being only *interpreted*.

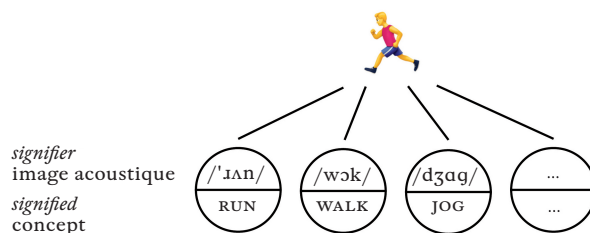


FIGURE 3. Example of the representation of an emoji within Saussure’s sign model

Another interesting example of emoji use can be found on the Twitter page of a Police Department located in the heart of St. Louis County

8. The second emoji is used probably in the same way as the sun emojis in Fig. 1.

(see <https://twitter.com/CreveCoeurPD>).⁹ On a regular basis, Creve Cœur Police post tweets in which security announcements are given concerning residents' properties and which have the goal of increasing public safety. These tweets imply that the use of emojis is no longer restricted to private everyday communication. However, it can still be assumed that the respective form of communication (in this case Twitter) continues to play an important role in how a text is structured; it is less likely that texts such as these tweets appear in print flyers, for example.

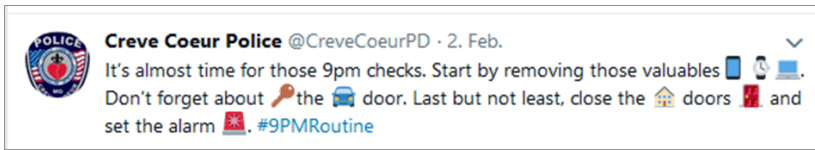


FIGURE 4. Emoji use in a non-private setting

In Fig. 4, the key emoji can be read as the verb *lock*, the car emoji as the noun *car*. The question arises as to how often emojis are used with such a logographic function. The data from the Swiss WhatsApp project suggest that this is actually only the case for a small number of all instances of emoji use. However, the text messages included in the corpus were collected in 2014; at that time, *Emoji Prediction* was not yet available to most writers. This software feature, introduced in 2014 on iOS and in 2016 on Windows Phone 8, facilitates the inclusion of emojis. The writer no longer needs to scroll through the list of emojis in order to find a suitable emoji, as context-related image suggestions are presented analogously to word suggestions. This also serves to highlight the substantial impact that technology exerts on writing.¹⁰ It is only due to the fact that emojis are available in the Unicode character set that we write with them today, and it is only because software presents emoji suggestions that they are increasingly functionalized logographically. This brings us to the next topic—the relation between Unicode and grapholinguistics.

9. Thanks to Marc Wilhelm Küster for bringing this to our attention. Cf. also Küster (in this volume).

10. The same was true for former SMS communication (cf. Bieswanger 2013). Certain writing strategies such as word junctions without spaces (*ShallWeMeetThisWeek-end?*) were established because of the 160 character limit of text messages.

4. Unicode and Grapholinguistics

In this section, the influence of the Unicode Consortium on writing will be discussed. In this context, it will also be argued that grapholinguists should have a say in the Unicode Consortium. Currently, the Consortium consists of about 20 people; Mark Davis co-founded it in 1991 and is its long-time president. All of the major IT companies (Apple, Microsoft, IBM, Facebook, Adobe, etc.) are “full members”. Additionally, there are three “institutional members” (e.g., the University of California in Berkeley) and two supporting members who also have the right to vote.¹¹ The members’ main task is to check applications for the admission of new characters and to make a preselection of these characters on which they vote in an annual meeting. The Consortium thus functions as a kind of gatekeeper (cf. Dürscheid 2018).

In the following, a short passage of the Unicode website is presented. This quote stresses the relevance of having a character coding system that facilitates the smooth exchange of data:

The Unicode Standard provides a unique number for every character, no matter what platform, device, application or language. It has been adopted by all modern software providers and now allows data to be transported through many different platforms, devices and applications without corruption. Support of Unicode forms the foundation for the representation of languages and symbols in all major operating systems, search engines, browsers, laptops, and smart phones—plus the Internet and World Wide Web [...].

<http://www.unicode.org/standard/WhatIsUnicode.html> <29.08.2019>

All Unicode characters (currently approx. 139,000) have a specific name (e.g., GREEK SMALL LETTER A) and are encoded with a numerical value. However, the concrete graphic realization that is finally assigned to a given Unicode code point and that appears on the device that is used to display the character depends on the specific font that is being used. Note, for example, how the respective sun emojis in Fig. 1 and 2 differ—even if just in details—with respect to their form and their color. They are concrete visual instantiations of the same basic shape. In this vein, from a grapholinguistic perspective, the vast majority¹² of Unicode characters are basic shapes that may be “embodied as graphs (sometimes referred to as *glyphs*), concrete physical instantiations” (Meletis, 2018, p. 63).

11. A complete list of the members is available at <http://www.unicode.org/consortium/members.html> <29.08.2019>.

12. There exist some Unicode characters which do not have a visual representation such as the soft hyphen character which marks boundaries between written syllables (cf. Haralambous and Dürst in this volume).

Unicode's predecessor was the ASCII character set ("American Standard Code for Information Interchange"). It originated in the 1960s, initially comprised only 64 characters and was eventually extended by one bit to 128. This, of course, led to various problems such as the faulty representation of characters from non-Roman scripts (such as Cyrillic). Moreover, many special characters from the Roman script could also not be represented correctly.¹³ Thus, for example, when sending emails including German umlauts or the sharp s (i.e., <ß>) in international correspondence, the German closing formula for *Best regards* could become *Sch%ne Grä#223;#223;e* (instead of *Schöne Grüße*) or instead of the <é> in the word *variété*, only an empty box could appear.

These times, however, belong to the past. In the long term, the goal of the Unicode Consortium is to integrate all scripts from the past and the present into the Unicode Standard. While the decision to include scripts currently in use appears self-evident, the question of why historical scripts should have a place in Unicode is justified. Consider, for example, a person wanting to write an article about cuneiform characters and to then publish it on the internet; without respective Unicode values, that person would have the option of inserting the cuneiform characters into the respective document as images. This is complicated and cumbersome; it is much easier and straightforward to type in the Unicode value of respective characters. Moreover, if image files were used, people who search the internet for articles on cuneiform characters would not find them with the aid of search engines. This goes to show that there are good reasons to include old scripts as well, which is how in Unicode, Egyptian hieroglyphics stand next to Germanic runes—to name just two examples. However, many historical scripts (e.g., Rongorongo) are still missing, as well as some scripts that are currently being used only by a small minority. These are listed on the website of the Script Encoding Initiative (SEI), a research project at the University of California at Berkeley (see below).

Obviously, every decision to include a new character in the Unicode Standard needs to be carefully examined since once a given character is added, it cannot be removed. This leads to the issue of the inclusion of emojis in Unicode and specifically the following question: Which criteria are crucial for an emoji coding proposal to be accepted or rejected? A page titled "How to Submit Proposal Documents" contains detailed information on this topic and lists points for and against accepting emoji proposals.¹⁴ For example, an emoji for a local food that is unknown in

13. The following passages are taken partially from an article that appeared in German under the title "Bild, Schrift, Unicode" (cf. Dürscheid 2018).

14. See http://unicode.org/emoji/proposals.html#selection_factors <30.09.2019>. Note that in July 2019, the Unicode Consortium launched a new website in celebration of the world emoji day (see <https://home.unicode.org/>

other regions (e.g., Swiss “Käsespätzle”) is rather unlikely to be included. Another important criterion is the assumed frequency of a prospective emoji: the emoji must represent something that is either in use world-wide or which is at least particularly frequent in a certain population group. Furthermore, an important criterion is whether it can be used in a sentence. This also explains why the emoji inventory contains so many characters that represent concrete things, such as sports equipment, means of transport, animals, and plants. If integrated into sentences (e.g., *I am 🍌, I love 🌺*), these for the most part culturally unspecific emojis are easy to decode for any reader.

An application for the inclusion of a new emoji can be submitted at any time. However, from application to final decision, up to two years can pass. Unsurprisingly, such a long-awaited decision is thus always expected with great excitement. Every year in June, the Unicode Consortium becomes the center of attention when it finally announces the new emojis to be introduced. Consider a small selection of headlines from the first half of 2019 (all accessed on 28.08.2019) which highlights the media’s and public’s interest in the introduction of new emojis:

New Emojis Are Coming: Interracial Couples, Guide Dogs, Falafel and More

<https://www.nytimes.com/2019/02/06/technology/new-emoji.html>

Disability emojis: Guide dog and wheelchair user revealed

<https://www.bbc.com/news/newsbeat-48989950>

One Woman Wants To Create This: *Insert Afro Emoji Here*

<https://www.npr.org/2019/03/31/708537582/one-woman-wants-to-create-this-insert-afro-emoji-here?t=1567008113435>

Unicode emoji 12.0: Waffles, otters and period positivity

<https://www.livemint.com/mint-lounge/features/unicode-emoji-12--0-waffles-otters-and-period-positivity-1550208051560.html>

For a long time, the work of the Consortium did not receive this kind of attention. It was, in fact, in 2010, precisely when emojis were included in the Unicode Standard, that this suddenly changed, which is also acknowledged on the Unicode website: “Emoji were adopted into the Unicode Standard in 2010 in a move that made the characters available everywhere. Today, emoji have been used by 92% of the world’s online population. And while emoji encoding and standardization make up just one small part of the Consortium’s text standards work, the growing popularity and demand for emoji have put the organization in the international spotlight.”¹⁵ This underscores not only Unicode’s importance

the-unicode-consortium-launches-new-website-in-celebration-of-world-emoji-day-2/<30.09.2019>). As noted in the press release, this website “will make information about the emoji proposal process more easily accessible while encouraging public participation and engagement in all Unicode initiatives”.

15. See [https://home.unicode.org/the-unicode-consortium-launches-new-website-in-celebration-of-world-emoji-day-2/<08.10.2019>.](https://home.unicode.org/the-unicode-consortium-launches-new-website-in-celebration-of-world-emoji-day-2/<08.10.2019>.”)

for global data exchange, but also that grapholinguists need to consider the consortium's work when investigating the impact of emojis on communication.

It is also worth noting that the decision to include emojis in the Unicode character set in the first place was certainly not an easy one. On the one hand, there were practical reasons in favor of their inclusion: emojis had already been used millions of times on Japanese mobile phones and large IT companies insisted on the need for globally uniform coding. On the other hand, the question arose whether emojis might only be a trend that would subside in a few years from then. Another question was whether images should be included in Unicode at all. And these questions only raise additional questions, including: Which criteria should be used in decision-making; which proposals should be accepted and which should be rejected? These are exactly the questions that Mark Davis, co-founder of the Unicode Consortium, addressed when he gave an interview in the Swiss newspaper *NZZ am Sonntag*.¹⁶ Every year, far more coding proposals are submitted than can be accepted, making a strict selection crucial. However, it is doubtful whether enough linguistic expertise is consulted when the consortium discusses these decisions.

This brings us to the point that is also advocated in Dürscheid's German publications. As mentioned above, the Unicode Consortium includes the representatives of all major internet companies (e.g., Adobe, Apple, Microsoft, Google) as full members and some additional institutional and supporting members. Among these, there is currently merely one researcher in linguistics: Dr. Deborah Anderson from the University of California. This should definitely change; linguists should have a lively interest in working on the future of the Unicode character set and the question of which basic shapes should be added (and which not). As for Deborah Anderson's background, she is a member of the Script Encoding Initiative (SEI), established in 2002, which is devoted to the preparation of proposals for the encoding of scripts in Unicode. As pointed out on its website, the SEI advocates the inclusion of minority and historic scripts into Unicode:

For a minority language, having its script included in the universal character set will help to promote native-language education, universal literacy, cultural preservation, and remove the linguistic barriers to participation in the technological advancements of computing. For historic scripts, it will serve to make communication easier, opening up the possibilities of online education, research, and publication.

<http://www.linguistics.berkeley.edu/sei/index.html> <30.09.2019>

16. See <https://nzzas.nzz.ch/gesellschaft/emojis-nachricht-mit-gefuehl-ld.1336511> <08.10.2019>.

Since today, in many literate societies, almost all reading and writing occurs digitally, it is essential to pay attention to the work of the Unicode Consortium, or even better: to participate in it. Linguists, and especially grapholinguists, must be actively involved in deciding what direction this process takes in the future. Not only do linguists have valuable insight into questions concerning the use of written language, but specialists in the field are also aware of the far-reaching sociolinguistic consequences of the introduction of digital writing in a given community.

5. Outlook

At the end of this paper, many questions remain unanswered: What will be the future role of the Unicode Consortium regarding the adoption of new characters? And what will be the future of emojis? Are they just a trend that will eventually disappear? In this vein, it must be noted that thanks to Unicode, it is to be expected that the number of emojis will increase continuously. However, it is also possible that new technologies will emerge that could make emojis obsolete. For example, voice messages might replace text messages and thus make emojis irrelevant. In any case, it will be interesting to observe how the relationship between image and writing will develop further and which graphics will still be used in the communication practices of the future. Moreover, it would be interesting to carry out another data collection of WhatsApp messages, comparing the new results with formerly described emoji practices in WhatsApp. Such a comparison of how writers employ emojis at various points in time might indicate that the frequency of emoji use is diachronically growing. And given the optimization of *Emoji Prediction*, emojis might also be increasingly used as logograms. Finally, it would be interesting to investigate whether emojis are on the rise also in contexts in which they were formerly not commonly used, for example in text genres such as business letters or on social media channels of universities, churches, museums, etc., i.e. in non-private settings. When considering the respective Instagram, Twitter and Facebook pages of such institutions, this already seems to be the case (see for the University of Zurich, for instance, <https://www.facebook.com/uzh.ch/>).

References

- Beißwenger, Michael and Steffen Pappert (2019). *Handeln mit Emojis. Grundriss einer Linguistik kleiner Bildzeichen in der WhatsApp-Kommunikation*. Duisburg, Essen: Universitätsverlag Rhein Ruhr.

- Bieswanger, Markus (2013). "Micro-Linguistic Structural Features of Computer-Mediated Communication". In: *Pragmatics of Computer-Mediated Communication*. Ed. by Susan C. Herring, Dieter Stein, and Tuija Virtanen. Berlin, Boston: de Gruyter, pp. 463–485.
- Dainas, Ashley R. and Susan C. Herring (in press). "Interpreting Emoji Pragmatics". In: *Internet Pragmatics: Theory and Practice*. Ed. by C. Xie, F. Yus, and H. Haberland. Amsterdam: John Benjamins.
- Danesi, Marcel (2016). *The Semiotics of Emoji. The Rise of Visual Language in the Age of the Internet*. London: Bloomsbury Publishing.
- Daniels, Peter T. (2018). *An Exploration of Writing*. Sheffield: Equinox.
- Dürscheid, Christa (2005). "Medien, Kommunikationsformen, kommunikative Gattungen". In: *Linguistik online* 22/1.
- (2016). *Einführung in die Schriftlinguistik*. 5th ed. Göttingen: Vandenhoeck & Ruprecht.
- (2018). "Bild, Schrift, Unicode". In: *Sprache—Mensch—Maschine. Beiträge zu Sprache und Sprachwissenschaft, Computerlinguistik und Informationstechnologie für Jürgen Rolsboven aus Anlass seines sechsundsechzigsten Geburtstages*. Ed. by Guido Mensching et al. Cologne: KUPS, pp. 269–285.
- Dürscheid, Christa and Karina Frick (2016). *Schreiben digital. Wie das Internet unsere Alltagskommunikation verändert*. Stuttgart: Kröner.
- Dürscheid, Christa and Christina M. Siever (2017). "Jenseits des Alphabets. Kommunikation mit Emojis". In: *Zeitschrift für Germanistische Linguistik* 45.2, pp. 256–285.
- Haralambous, Yannis and Martin Dürst (in this volume). "Unicode from a Linguistic Point of View".
- Herring, Susan C. and Ashley R. Dainas (2017). "Nice Picture Comment! Graphicons in Facebook Comment Threads". In: *Proceedings of the Fiftieth Hawaii International Conference on System Sciences (HICSS-50)*. Los Alamitos: IEEE Press, pp. 2185–2194.
- Herring, Susan C., Dieter Stein, and Tuija Virtanen, eds. (2013). *Pragmatics of Computer-Mediated Communication*. Berlin, Boston: de Gruyter.
- Jucker, Andreas H., Heiko Hausendorf, Christa Dürscheid, et al. (2018). "Doing Space in Face-to-Face-Interaction and on Interactive Multimodal Platforms". In: *Journal of Pragmatics* 134, pp. 85–101.
- Küster, Marc Wilhelm (in this volume). "Open and Closed Writing Systems. Some Reflections".
- Meletis, Dimitrios (in press). "The Grapheme as a Universal Basic Unit of Writing". In: *Writing Systems Research*.
- (2018). "What is Natural in Writing? Prolegomena to a Natural Grapholinguistics". In: *Written Language and Literacy* 21.1, pp. 52–88.
- (2019). "Naturalness in Scripts and Writing Systems: Outlining a Natural Grapholinguistics". PhD thesis. University of Graz.

- Neef, Martin (2015). “Writing Systems as Modular Objects: Proposals for Theory Design in Grapholinguistics”. In: *Open Linguistics* 1, pp. 708–721.
- 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.
- Ueberwasser, Simone and Elisabeth Stark (2017). “What’s Up, Switzerland? A Corpus-Based Research Project in Multilingual Switzerland”. In: *Linguistik Online* 84/5. <https://bop.unibe.ch/linguistik-online/article/view/3849/5833> <06.09.2019>.
- Unger, J. Marshall (1990). “The Very Idea: The Notion of Ideogram in China and Japan”. In: *Monumenta Nipponica* 45.4, pp. 391–411.
- Weingarten, Rüdiger (2011). “Comparative Graphematics”. In: *Written Language and Literacy* 14.1, pp. 12–38.

What Do Kanji Graphs Represent in the Current Japanese Writing system? Towards a Unified Model of Kanji as Written Signs

Keisuke Honda

Abstract. In the current Japanese writing system, kanji graphs constitute a major subpart of its signary. There are two opposing views on how to characterise them in linguistic terms, making different claims about the type of linguistic unit they represent. The first view claims that kanji graphs are based primarily on the morpheme because a majority of currently used graphs represent individual morphemes. The second view maintains that they are based primarily on the sound and only secondarily on the morpheme because all graphs represent sounds that may or may not correspond to individual morphemes. The present paper discusses the advantages and disadvantages of both views and sketches out a new, unified model of how kanji graphs function as written signs. In this model, kanji graphs are seen as the formal building blocks of simplex or complex written signs representing the phonological exponents of individual morphemes.

Introduction

This paper discusses the type of linguistic unit represented by the graphs of the kanji (漢字) script in the present-day Japanese writing system.

The starting point of the present article is a practice widely observed in graphemics, in which individual writing systems are referred to as being ‘phonemic,’ ‘syllabic,’ ‘morphemic’ and so on (Section 1). More specifically, each writing system is characterised in terms of a particular type of linguistic unit (e.g., phoneme) if all or most of its written signs represent the individual instances of that unit (e.g., /i/, /a/, /o/, /p/, /t/, /k/, ...). A prerequisite for such a characterisation is a linguistic analysis of the signary, that is, the set of written signs employed in the given writing system. The validity of the characterisation, then, depends on the adequacy of the signary analysis.

Keisuke Honda  0000-0003-4228-5406

Imperial College London, Centre for Languages, Culture and Communication
South Kensington Campus, London SW7 2AZ, United Kingdom

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

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1.
Fluxus Editions, Brest, 2019, p. 185–208. <https://doi.org/10.36824/2018-graf-hond>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

In this regard, Japanese kanji graphs deserve special attention (Section 2). They are sinographs or ‘Chinese characters’ employed in the Japanese writing system, used alongside the graphs of the kanji-derived hiragana (平仮名) and katakana (片仮名) scripts as well as the Latin script known as rōmaji (ローマ字) (Smith 1996, pp. 209–210; Honda 2012, pp. 39–47; I. Taylor and M. M. Taylor 2014, pp. 271–283).¹ Despite the persisting belief that kanji graphs represent things and ideas without recourse to language (e.g., Suzuki 1975, p. 178), they are in fact closely related to the phonological, morphological and semantic properties of the Japanese lexicon (e.g., Unger 1987, pp. 45–49, 1990, pp. 397–411; Kōno 1994, p. 11; Matsunaga 1996, pp. 2–12). Today kanji graphs are used mainly to write individual content words or their stems in the Sino-Japanese, native Japanese and hybrid vocabularies. Given the mixed use of kanji and other graphs, kanji graphs constitute what may be seen as a major subpart of the signary of the Japanese writing system. An important question, then, is whether a single type of linguistic unit should be postulated to account for the functioning of kanji graphs and, if so, what that unit might be.

In the literature, it is possible to identify two major schools of thought on this question (Sections 3 and 4). For convenience, the present paper refers to the first one as the *morphographic theory* and the second one as the *morphophonetic theory*, borrowing the respective terms from Joyce (2011, p. 58) and Matsunaga (1996, p. 17).² The morphographic theory claims that kanji should be considered primarily morphemic because there is a one-to-one correspondence between individual graphs and morphemes in most kanji-written words (Hill 1967, pp. 93–96; Miller 1967, pp. 92–93; 1986, 15ff; Nomura 1999, pp. 1–3; Sproat 2000, pp. 154–160; Joyce 2001, pp. 12–111; 2011, pp. 63–72; Sampson 2015, pp. 208–232). In contrast, the morphophonetic theory holds that kanji graphs are primarily phonographic and only secondarily morphemic because all graphs represent sounds that may or may not correspond to individual morphemes (DeFrancis 1989, pp. 138–143; Matsunaga 1994, pp. 34–39, 1996, pp. 14–18; also see Unger 1987, pp. 35–49; DeFrancis and Unger 1994; Unger and DeFrancis 1995).

To the knowledge of the present author, there has been little attempt to examine the validity of these two theories through direct comparison. However, they deserve special consideration because they provide

1. For a comprehensive description of kanji graphs and their use, see Satō (1987–1989), Satō et al. (1996), and Kōno, Nagata, and Sasahara (2001), among others.

2. DeFrancis (1989, p. 58) first proposed the term ‘morphophonetic,’ together with the alternative form ‘morphonic’. The present paper adopts the former, although with a warning not to confuse it with the unrelated term ‘morphophonemic,’ because the latter is conventionally associated with the notion of ‘morphon’ in Stratificational Grammar (Lamb, 1966).

significantly different interpretations of the way kanji graphs function as written signs. According to the morphographic theory, kanji graphs relate directly to the morphological level of linguistic representation. An important implication of this notion is that they function in a fundamentally different way from phonographs or phonologically based written signs. Contrastively, the morphophonetic theory suggests that kanji graphs relate mainly to phonology and only optionally to morphology. This implies that they share a common ground with phonographs, in that sounds play a crucial role in both types of written signs. These considerations motivate a comparative examination of the two existing theories.

This paper presents a critical analysis of the morphographic and morphophonetic theories and develops a preliminary sketch of a new, unified model of how kanji function as written signs in the current Japanese writing system. Section 1 introduces the notion of linguistic unit underlying a writing system. Section 2 provides the necessary background on kanji graphs and kanji-written words. Section 3 takes a closer look at the morphographic theory, with a particular focus on the analysis of two-kanji compound words presented by Joyce (2001; 2011). Section 4 turns to the morphophonetic theory, focusing on Matsunaga's (1994; 1996) discussion of what are known as the phonetic elements of kanji graphs. Section 5 proposes an integrated model of kanji as written signs, which draws on the advantages of the existing theories while avoiding their disadvantages. In this proposed model, kanji graphs are viewed as the formal building blocks of structurally simplex or complex written signs representing the phonological exponents of individual morphemes. Section 5 summarises the discussion and draws conclusions.

1. Linguistic Unit Underlying a Writing System

Writing may be seen as a system of visible and/or tactile marks, by means of which utterances can be encoded into and decoded from particular graphical representations in a more or less conventional manner (Daniels 1996, p. 3; 2018, pp. 156–157; Coulmas 2003, pp. 1–17). This paper refers to such marks individually as *graphs* (Sampson, 2015, pp. 10–11) and collectively as a *script* (Sproat, 2000, p. 25). Graphs may be used individually (e.g., <p>) or in fixed combinations (e.g., <pp>) to represent distinct sounds (e.g., /p/), sound combinations (e.g., /pa/), or sound-meaning units (e.g., /papa/ 'father'). Thus, one may speak of *written signs*, each formed by an arbitrary association of a graphical form as the signifier (e.g., <p>) and a linguistic value as the signified (e.g., /p/).³ When a

3. This account is based on Saussure's (1916) dyadic model of signs. It remains an open question whether this model is in any way preferable to Peirce's (1931–1958)

set of written signs is used in accordance with a body of conventions to write a particular language, it is common to regard them respectively as the *signary* and the *orthography* of a *writing system* (Daniels and Bright 1996, pp. xliii–xliv; Coulmas 2003, pp. 35–36).

The notion of *linguistic unit* plays a crucial role in graphemics or the linguistic study of writing systems. Firstly, it is commonly assumed that each written sign represents a specific instance of a particular linguistic unit. For example, a written sign is said to represent a phoneme if its graph corresponds to a single vowel or consonant (e.g., ‘Finnish <p> represents /p/’), or a morpheme if the graph has a sound-meaning value (e.g., ‘Chinese 山 represents {mountain} = /shān/ ‘mountain’) (for more examples, see descriptions of the world’s writing systems in Daniels and Bright 1996 and Kōno, Chino, and Nishida 2001). Secondly, as already mentioned in the introduction to this paper, it is common to characterise a writing system in terms of a particular linguistic unit (e.g., ‘Finnish is a phonemic writing system,’ ‘Chinese writing system is morphemic’) (again, see Daniels and Bright 1996 and Kōno, Chino, and Nishida 2001). The underlying assumption seems to be that every writing system can be—or even should be—described in terms of one single type of linguistic unit that is most relevant to the signary of that system.⁴

Gelb (1963, pp. 190–205) provides a clear formulation of this assumption in his explanation of what he terms the ‘evolution of writing’. Gelb notes that “[t]here are no pure systems of writing” because any writing system “may contain elements from different phases of its development” (pp. 199–200). To cite his examples, the English writing system employs some word signs (e.g., <£>) in addition to phonemic signs (e.g., <p>) (p. 200).⁵ Nonetheless, Gelb describes English as being ‘alphabetic’. In other words, the entire writing system is characterised as being phonemic, abstracting away from the use of word signs. This, according to

triadic model of signs for a better understanding of writing (Gerald Penn, personal communication, 14th June 2018).

4. Sometimes compound descriptors like ‘morphosyllabic’ are also used (e.g., DeFrancis 1989, p. 58; see Section 4.1). However, assuming a headed structure in such compounds, it is reasonable to interpret the head as the main part of the characterisation (i.e., *morphosyllabic*). Moreover, at least in English writing, hyphenation would be used if they are meant to be *dvandva* compounds (i.e., *morpho-syllabic*). Thus, as far as English is concerned, ‘morphosyllabic’ should be interpreted as ‘primarily syllabic’ (Kaiser, 1995, p. 163).

5. This currency symbol can be seen as a word sign because it represents a particular sound-meaning unit (i.e., <£> /paʊnd/ ‘currency unit’) rather than a sound sequence (e.g., not *<com£> for *compound*). However, it can also be interpreted as a morpheme sign when it is used for writing the monomorphemic *pound* as in <£1>, and as a word sign when used for the polymorphemic *pound+s* as in <£2>. It may also be considered as being ideographic when reduplicated as in <£££>, rendered variously as *hundreds of pounds*, *three-digit pounds*, *a lot of money* and so on.

Gelb, is justified on the grounds that it allows one to capture “only the major characteristics” of the writing system (p. 200).

If one accepts Gelb’s (ibid.) above observation that there are no pure writing systems, it would follow that the signary of every writing system contains different subsets of signs based on distinct types of linguistic units. In that case, characterising a given writing system in terms of a single linguistic unit would presuppose a distinction between units of primary and secondary importance. By calling a given system alphabetic or phonemic, for instance, one is implicitly or explicitly stating that the phoneme is central—and all other units peripheral—to the functioning of its signary. Such a distinction must be made on the basis of a thorough and systematic linguistic analysis of the signary. This point deserves emphasis because an inadequate analysis could lead to an inaccurate description of the writing system in question.

2. Kanji Graphs and Kanji-Written Words

One of the most striking features of the current Japanese writing system is its mixed use of multiple scripts (Backhouse 1984, p. 219; Smith 1996, p. 214; Joyce 2001, p. 12; Joyce 2011, p. 62; Honda 2012, pp. 38–39). As already mentioned above, there are four main scripts currently in use, namely the sinographic kanji, kanji-derived hiragana and katakana, and the Latin script known as rōmaji. While it is theoretically possible to write Japanese entirely in one of these scripts, the norm is to use all of them for different purposes in a complementary way.⁶ In other words, the four scripts function as distinct but interlinked subparts of a complex signary in the current Japanese writing system.

The kanji script constitutes the largest of those subparts. Currently some 2,000 to 3,000 kanji graphs are in common use, together with another few thousand graphs of relatively low frequency (Joyce, 2001, pp. 17–19). A majority of these graphs were historically imported from the Chinese writing system, while others were invented in Japan following the same formation principles underlying the imported ones (Satō

6. This functional division, which is non-binding but commonly observed, can be outlined as follows: (1) kanji graphs are used for content words and morphemes (see below); (2) hiragana graphs are used for grammatical particles, derivational and inflectional affixes, as well as some content words; (3) katakana graphs are used for modern loanwords, native mimetics and the names of flora and fauna; and (4) rōmaji graphs are used for foreign words and abbreviations of native and non-native words. Some might oppose the possibility of writing Japanese solely in kanji graphs, saying that they cannot indicate grammatical information. However, this is a viable option in view of the historical use of *man'yōgana* (万葉仮名) or phonographically employed kanji graphs (e.g., Seeley 2000, p. 190).

1987–1989; Seeley 2000; Frellesvig 2010; Okimori 2011). The Chinese-made graphs were initially adopted to read and write texts in classical Chinese. They were gradually adapted to write both what had become Sino-Japanese (SJ) lexical items and their native Japanese (NJ) equivalents by way of translation. On the other hand, the Japanese-made graphs, known as *kokuji* (国字) or ‘national characters,’ were used to write NJ lexical items that had no equivalents in Chinese.

Today both classes of kanji graphs are used to write a large subset of Japanese lexical items, which are etymologically SJ (e.g., 書物 /shomotsu/ ‘book’), NJ (e.g., 書留 /kakitome/ ‘registered post’) or a hybrid of both (e.g., 書棚 /shodana/ ‘bookshelf’).⁷ Although some content words are written with individual graphs (e.g., 書 /sho/ ‘writings’), the majority are written with strings of two or more graphs (e.g., 書道 /shodō/ ‘calligraphy,’ 書道家 /shodōka/ ‘calligrapher’). *Kanji* graphs may also be combined with hiragana graphs to write inflected words (e.g., 書く /kaku/ ‘write’) and derived forms thereof (e.g., 書き /kaki/ ‘the way one writes something’), as well as a small number of non-inflected words (e.g., 且つ /katsu/ ‘besides’).⁸ They may also be used in combination with hiragana or katakana graphs to write hybrid compounds (e.g., 書道セツト /shodōsetto/ ‘set of calligraphy tools’).

As noted in Section 1, a written sign can be seen as an arbitrary association of a graphical form and a linguistic value. Assuming that kanji graphs constitute the forms of written signs, it is possible to isolate their values through a comparative analysis of kanji-written words. For instance, a comparison of such words as 書 /sho/ ‘writings,’ 書物 /shomotsu/ ‘book’ and 書棚 /shodana/ ‘bookshelf’ reveals that the graph 書 has the value /sho/, which conveys ‘writing’ and other related meanings. Traditionally, the value of a kanji graph is referred to as *yomi* (読み) or, in English, ‘readings’. Each reading consists of a particular pronunciation which often, but not always, denotes a specific meaning (Section 3.2). Due to the historical background of kanji graphs and kanji-written words described above, a single graph may be associated with an *on* (音) or SJ reading, a *kun* (訓) or NJ reading, or both.⁹ It is also common

7. *Kanji* graphs may also be used to write non-Chinese loanwords (e.g., 煙草 *tabako* ‘tobacco,’ 浪漫 *roman* ‘romanticism’). However, this usage is confined to a small subset of the vocabulary and is often replaced by hiragana or katakana writing (e.g., kanji 煙草 by katakana タバコ).

8. In this usage, there is often a mismatch between the kanji-hiragana boundary and the morpheme boundary within a word. To illustrate with 書く /kaku/ ‘write’ (morphologically *kak-u* ‘write-non.past-aff-plain’), the hiragana く corresponds to both the stem-final /k/ and the suffix /u/. In the literature, there are different approaches to account for such a mismatch (e.g., Kaiser 1995, p. 165; Honda 2012, pp. 133–142). The present paper leaves this topic for future research.

9. With regard to the types of readings, the present paper uses the terms *on* and *kun* instead of SJ and NJ. This is because some readings commonly thought to be NJ

for a single graph to have multiple *on* readings, multiple *kun* readings or both, owing to the fact that kanji-written words were borrowed from different dialects of Chinese, and then translated into Japanese by different schools of literate traditions. For example, the graph 音 has two *on* readings /on/ and /in/, and two *kun* readings /oto/ and /ne/, all meaning ‘sound’.

There are two special uses of kanji graphs which require particular mention here. The first one is *jukuji* (熟字) or ‘polygraphic character,’ in which a string of two or more graphs forms a single functional unit and corresponds to a lexical item in a many-to-one manner. Used this way, the graph string is said to carry a special *kun* reading known as *jukujikun* (熟字訓), sometimes translated as ‘idiomatic *kun*’ (I. Taylor and M. M. Taylor, 2014, p. 279). One example of *jukuji* is 田舎, which has the *jukujikun* /inaka/ ‘countryside’. Importantly, the graph 田 is usually rendered in the *on* reading /den/ or the *kun* reading /ta/, both meaning ‘rice field,’ and 舎 in the *on* reading /sha/, meaning ‘hut’. As this example illustrates, a *jukujikun* is not the total sum of the regular readings of the graphs constituting the given *jukuji*.

The second special use of kanji graphs includes *on’yaku* (音訳) and *ateji* (当て字), both involving what is known in the literature as the ‘rebus principle’ (Coulmas, 1996, pp. 433–434). *On’yaku*, which may be translated as ‘phonetic translation,’ was historically used to transcribe non-Chinese loanwords like 檀那 /danna/ ‘master’ (< Sanskrit *dāna*) and 襦袢 /juban/ ‘undershirt’ (< Portuguese *gibão*) (NKDDHI, 2000–2002). In both examples, each kanji graph is used for the phonological property of its regular reading without regard to the meaning. To illustrate this point, the readings of both 檀 /dan/ ‘cedar, sandalwood’ and 那 /na/ ‘that, which’ are used purely phonologically in the first example above, abstracting from their etymologically irrelevant meanings. The same principle underlies *ateji*, roughly translated as ‘assigned character,’ which refers to rebus notation of non-Chinese loans such as 浪漫 /roman/ ‘romanticism’ as well as NJ lexical items like 野暮 /yabo/ ‘unrefined’ (ibid.).

Finally, a special mention should be made of the *Jōyō Kanji Hyō* (常用漢字表) or ‘List of Characters for General Use’ (Japanese Cabinet, 2010). This is a body of guidelines on the use of kanji graphs and their readings, defined for everyday purposes by the Japanese Ministry of Education. First promulgated by the Japanese Cabinet in 1981, the list went through a partial revision, and a new version was issued in 2010. The current list contains 2,136 graphs and 4,388 readings (2,352 *on* and 2,036 *kun*), together with examples of common words written with them. Although legally non-binding, these graphs and readings are widely accepted as

in fact originate in Chinese (e.g., 馬 /uma/ ‘horse’) or Korean (e.g., 寺 /tera/ ‘temple’) (NKDDHI, 2000–2002).

a de facto standard for kanji orthography.¹⁰ Nevertheless, it should be stressed that these graphs and readings constitute only a subset of those actually used in the current Japanese writing system. In this sense, the List of Characters for General Use must be seen as a representative sample and not as the whole picture of kanji usage.

3. The Morphographic Theory

Turning now to the main subject of this paper, the morphographic theory sees the morpheme as the primary linguistic unit underlying the functioning of kanji graphs. While there are some different ways to define what a morpheme is, a textbook definition is that it is “the smallest unit of language that carries information about meaning or function” (O’Grady and de Guzman, 1997, p. 133). A morpheme can form a word by itself, in which case the word in question is said to be *monomorphemic* or *morphologically simplex*. It can also be concatenated with another morpheme to form a *polymorphemic* or *morphologically complex* word. In English, for example, the morpheme {write} can stand by itself as the monomorphemic word *write*, or form a part of polymorphemic words like *writing* and *writer*. With the notion of morpheme in mind, this section takes a close look into the morphographic theory of kanji graphs.

3.1. An Overview of the Morphographic Theory

The term *morphography*, also known as *morphemic writing*, refers to a one-to-one correspondence between graphs and morphemes (e.g., Joyce 2011, p. 59; Sampson 2015, 23ff).¹¹ As already introduced above, the morphographic theory holds that such a correspondence can be observed across kanji-written words. This theory is accepted by many studies in the field of Japanese linguistics, which describe kanji graphs as ‘morphemic writing’ (e.g., Miller 1967, 92–93ff, 1986, 15ff) or *byōkeitaiso moji* (表形態素文

10. The 2,136 kanji graphs account for over 96% of all tokens of kanji-written words found in the 100-million word Balanced Corpus of Contemporary Written Japanese (Joyce, Masuda, and Ogawa, 2014, pp. 177–178).

11. Morphography differs from *logography* or the representation of individual words, and *phonography* or the representation of phonological units such as phonemes or syllables. While various instances of morphography can be found in the world’s writing systems (Daniels and Bright 1996; Kōno, Chino, and Nishida 2001), views differ on whether it is possible to develop full-fledged writing based entirely or primarily on morphography (e.g., Hill 1967; DeFrancis and Unger 1994; Sproat 2000; Sampson 2015).

字), roughly translated as ‘morpheme-representing characters’ (e.g., Nomura 1999, pp. 1–3). It is also widely endorsed in general writing systems research, where kanji graphs are commonly characterised as a morphographic component of the Japanese writing system (e.g., Hill 1967, pp. 93–96; Sproat 2000, pp. 154–160; Sampson 2015, pp. 208–232). In this context, Joyce (2001, pp. 12–111; 2011) deserves particular attention because his study offers a powerful empirical basis for examining the morphographic theory.

At the heart of Joyce’s (2001; 2011) discussion is the notion of *morphographic principle*, which he claims is fundamental to the way kanji graphs function. Under this principle, individual graphs not only represent morphemes but are also spatially arranged in accordance with the morphological structure of the word being written. Joyce maintains that this is the case in a vast majority of kanji-written words. To support this, he presents a morphological analysis of two-kanji compound words, that is, Japanese words written by combining two separate kanji graphs. They include SJ and NJ words as well as their hybrids, which, according to a dictionary-based survey of kanji-written words cited by Joyce, account for up to 70 per cent of all Japanese words (Yokosawa and Umeda, 1988, p. 377). Based on Nomura’s (1988a; 1988b) study of word-formation patterns in kanji-written words, Joyce distinguishes nine principles underlying two-kanji compound words. These are presented in Table 1 below, reproduced with the original examples from Joyce (2011, p. 71, Table 3). For each principle, the left column shows two glossed examples and the right column indicates whether the principle in question is morphologically motivated or not.

According to Joyce (2001; 2011), the first eight principles are morphologically motivated, meaning that they involve the concatenation of two morphemes (e.g., 国道 /kokudō/ ‘national road’ = {country} + {road}). In writing, kanji graphs correspond to these morphemes and are linearly arranged in the same order as they are concatenated (e.g., 国 {country} + 道 {road}). Joyce maintains that the only non-morphologically motivated principle is the last one, designated as ‘phonetic borrowing’. In his terminology, this is an umbrella term for words written in *jukuji*, *on’yaku* or *ateji* (Section 2). Individual kanji graphs do not correspond to morphemes either in *jukuji*, where they constitute polygraphs, or in *on’yaku* and *ateji*, where they function phonographically. Joyce dismisses words formed by phonetic borrowing as being “by far the exception” (Joyce, 2011, p. 71) to the predominantly morphological nature of two-kanji compound words and, by extension, the principally morphographic nature of kanji graphs. This is justified on the basis of Gelb’s (1963, p. 199) above-mentioned observation that there are no pure writing systems. Thus, Joyce sees the morpheme as the primary linguistic unit underlying the functioning of kanji graphs.

TABLE 1. Word-formation principles underlying two-kanji compound words (reproduced from Joyce 2011, p. 71)

<i>Principle</i>			<i>Morphological</i>
Modifier + modified			Yes
山桜 /yamazakura/	‘mountain’ + ‘cherry’	= mountain cherry	
国道 /kokudō/	‘country’ + ‘road’	= national road	
Verb + complement			Yes
登山 /tozan/	‘climb’ + ‘mountain’	= mountain climbing	
殺人 /satsujin/	‘kill’ + ‘person’	= murder	
Complement + verb			Yes
外食 /gaishoku/	‘outside’ + ‘eat’	= eat out	
毒殺 /dokusatsu/	‘poison’ + ‘kill’	= kill by poison	
Associative pairs			Yes
親子 /oyako/	‘parent’ + ‘child’	= parent(s) and child(ren)	
生死 /seishi/	‘life’ + ‘death’	= life and death	
Synonymous pairs			Yes
山岳 /sangaku/	‘mountain’ + ‘mountain’	= mountains	
変化 /henka/	‘change’ + ‘change’	= change	
Repetitions			Yes
段々 /dandan/	‘step’ + ‘step’	= gradually, by degrees	
個々 /koko/	‘piece’ + ‘piece’	= individual, one by one	
Derivation			Yes
不明 /fumei/	‘un-’ + ‘clear’	= unclear, obscure	
史的 /shiteki/	‘history’ + ‘-ic’	= historic	
Abbreviations			Yes
農協 /nōkyō/	from 農業共同	= agricultural cooperative	
春闘 /shuntō/	from 春季闘争	= spring (labor) offensive	
Phonetic borrowing			No
葡萄 /budō/		= grapes	
面倒 /mendō/		= care	

3.2. Problems of the Morphographic Theory

Joyce’s (2001; 2011) analysis of two-kanji compound words provides a strong empirical basis for the morphographic theory of kanji graphs. At the same time, it faces at least two major problems that have gained little attention in the literature.

3.2.1. *Semantic Transparency*

The first problem is best captured by making reference to the notion of *semantic transparency* or the extent to which the meaning of a polymorphic word can be predicted from the meanings of its constituent morphemes (Körtvélyessy, Štekauer, and Zimmermann, 2015, pp. 87–92). It is generally conceived as a scalar notion (i.e., greater-or-lesser) rather than a binary one (i.e., either-or), meaning that a given word may be considered more or less transparent than another.

Two intertwined factors contribute to semantic transparency, namely *compositionality* and the presence of *constant meanings* in word elements. To exemplify with the English word *blueberry*, it is analysable into two meaningful elements *blue* and *berry* through comparison with other words like *bluebird* and *blackberry*. Because these elements are not further analysable into smaller meaningful parts, they can be considered as two separate morphemes. Besides, one may speak of a part-to-whole relationship between the meanings of these morphemes (i.e., ‘a colour’ and ‘a small roundish fruit’) and that of the compound they constitute (i.e., ‘a berry of that colour’). In this sense, *blueberry* can be seen as a semantically transparent compound of {blue} and {berry}. In contrast, semantic transparency is less evident in *strawberry*. While this word is also analysable into {straw} and {berry}, the meaning of the first morpheme (i.e., ‘stalk of a cereal plant’) is less clearly related to that of the compound when compared to that of {blue} in *blueberry*. This is even more so in *cranberry*, as the element *cran-* occurs only in this particular word and its meaning is therefore unidentifiable by way of comparison.

The notion of semantic transparency poses a serious challenge to Joyce’s (2001; 2011) analysis of two-kanji compound words, which is pivotal to his argument for the morphographic theory. As noted above, Joyce assumes morphological constituency in most types of two-kanji compound words, with the sole exception of those formed by phonetic borrowing. This assumption predicts compositionality in such words because the presence of constant meanings is a prerequisite for the analysis of words into morphemes. To borrow an example from Table 1 above, Joyce (2011) categorises the commonly used word 変化 /henka/ ‘change’ as a synonymous pair and analyses it into 変 /hen/ ‘change’ and 化 /ka/ ‘change’. This analysis is plausible in view of words like 変心 /henshin/ ‘change of mind’ and 化成 /kasei/ ‘transformation’. Given the clear relationship between the meanings of the word elements and that of the compound itself, it seems reasonable to assume a certain degree of compositionality and, by extension, a morphological constituency in this word. As Vance (2002, p. 187) points out, however, it is often dubious to assume a similar degree of compositionality in words like 勉強 /benkyō/ ‘study’. Also a common word formed by synonymous pair, it is analysable into 勉 /ben/ ‘striving’ and 強 /kyō/ ‘strength’ through

comparison with items like 勉勵 /benrei/ ‘diligence’ and 強風 /kyōfū/ ‘strong wind’. Nevertheless, unlike the straightforward compositionality in 變化 /henka/ ‘change,’ it is not immediately clear why the combination of ‘striving’ and ‘strength’ results in 勉強 /benkyō/ ‘study’.¹² In this light, it appears plausible to say that the degree of compositionality is higher in the first example and lower in the second one. This observation calls into question the notion of morphological constituency as an essential feature of all two-kanji compound words except for those formed by phonetic borrowing.

A key factor overlooked by Joyce (2001; 2011)—and in fact also by many proponents of the morphographic theory—is diachronic changes in lexical meanings. As for 勉強 /benkyō/ ‘study,’ there is evidence that this word underwent a semantic shift from the original meaning of ‘diligence’ to the current meaning of ‘study’.¹³ Although rather impressionistic, the meanings of 勉 /ben/ ‘striving’ and 強 /kyō/ ‘strength’ appear to be more closely related to this former meaning than to the latter one. If one accepts this interpretation, then it would be possible to say that the word under discussion has become less compositional over the course of history. As a matter of fact, such a decrease in compositionality can be observed in many two-kanji compound words. Of particular importance are words containing kanji graphs with obsolete meanings (Nomura 1999, p. 10; Tajima 2006, pp. 6–8). One example is 挨拶 /aisatsu/ ‘greeting,’ another commonly used synonymous pair word. Historically, it was a compound of 挨 /ai/ ‘push’ and 拶 /satsu/ ‘shove,’ denoting a religious practice of Zen Buddhism in which a monk would ‘press’ his peer verbally or even physically to test his level of enlightenment (NKDDHI, 2000–2002). At present, however, this meaning has become obsolete and can only be confirmed by consulting dictionaries and other reference resources. It is also important to note that the graphs 挨 and 拶 normally occur only in this particular combination.¹⁴ Consequently, there is no way to isolate their present-day meanings—if they existed—by means

12. One might suspect that this is due to the English translations of the original meanings provided here. However, the situation remains by far the same even in view of other translations. For instance, Nelson’s (1997) *Japanese-English Character Dictionary* gives the following translations: 勉 ‘serve, fill a post, serve under; exert oneself, endeavour, work, be diligent; play (the part of); as much as possible; diligently’; 強 ‘strength, might; strong person’.

13. This original meaning is attested in *Mōshibō* (毛詩抄), a collection of lecture notes compiled in the first half of the 17th century, whereas the current meaning probably came about in the 19th century (NKDDHI, 2000–2002).

14. One exception is the variant form 一挨一拶 /ichiaiiissatsu/ ‘one pushing, one shoving,’ which denotes the same Zen practice described above. In historical usage, 挨 and 拶 also occur in combination with other graphs, as instantiated by 挨次 /aiji/ ‘consecutive’ and 逼拶 /hissatsu/ ‘put pressure’. However, there are only a handful of such words (ibid.).

of comparison. Therefore, as far as the contemporary Japanese lexicon is concerned, it is safe to conclude that 挨拶 /*aisatsu*/ ‘greeting’ has lost its historical compositionality and, hence, morphological constituency.¹⁵

One might be tempted to tackle this problem by attaching lesser importance to the role of meaning in morphemehood. As already noted, a textbook definition of morpheme is that it is the smallest linguistic unit carrying information about meaning or function. For kanji-written words, however, Miyajima (1973, p. 15) postulates a special kind of morpheme called *muimi keitaiso* (無意味形態素) or ‘meaningless morpheme’. It is defined as “an element carrying no active meaning by itself, which always occurs in combination with certain other (meaningful) elements” (English translation by the present author).¹⁶ Following Bloomfield (1933, 160ff), he equates meaningless morpheme to the *cran-*element in English *cranberry* in that it carries no denotational meaning but a differential meaning (i.e., standing for nothing but distinguishing *cranberry* from *blackberry*, *strawberry*, *gooseberry*, etc.). If one accepts this notion, it might be possible to treat kanji graphs like 挨 and 拶 as representing meaningless morphemes. However, such a treatment would obfuscate the delineation of morpheme and call for a radical reconceptualisation of morphography.

3.2.2. Orthographic Variation

The second problem concerns synchronic and diachronic variation in the orthographic forms of two-kanji compound words. Synchronically, there are a number of two-kanji compound words in which a kanji graph can be replaced with another one without changing the word’s meaning. To give one example, both 少食 and 小食 are commonly used to write /*shōshoku*/ ‘light eating’. Both 少 and 小 are associated with the *on* reading /*shō*/, which means ‘few, little’ in the former and ‘small’ in the latter. Assuming the traditional definition of morpheme as the smallest meaningful unit, they must be treated as representing homophonous but semantically related morphemes. This treatment is faced with the additional task of proving that 少食 /*shōshoku*/ and 小食 /*shōshoku*/ are distinct words denoting different meanings (e.g., ‘eating little amount of

15. Morioka (2004, p. 102) reports that there are approximately 950 kanji graphs with obsolete meanings like 挨 and 拶 within the set of 6,355 common kanji graphs defined by the Japanese Industrial Standard for IT use. These include graphs used for writing common words (e.g., 絢爛 /*kenran*/ ‘gorgeous,’ 狡猾 /*kōkatsu*/ ‘cunning’) as well as those for relatively infrequent ones (e.g., 踟躕 /*kyokuseki*/ ‘cower,’ 魍魎 /*mōryō*/ ‘spirits and goblins’).

16. The original definition reads as follows: “それ自身では積極的な意味をもっておらず、つねにほかの特定の(有意味的な)要素と結びついてあらわれる要素” (Miyajima, 1973, p. 15).

food' versus 'eating small size food'). In reality, however, nothing seems to suggest that this is the case. Alternatively, one might argue that 少 and 小 represent two meaningless morphemes, but this argument is also untenable because the readings of these graphs are clearly distinguished in terms of meaning (i.e., 'few, little' versus 'small').

The situation becomes further complicated if diachronic variation is also taken into account. One interesting example is the common word 時計 /tokei/ 'timepiece' (NKDDHI 2000–2002; Tajima 2006, pp. 11–12). Superficially, it seems analysable into 時 /toki/ 'time' and 計 /kei/ 'measure,' which, despite the phonological discordance, might appear semantically transparent to some degree. Before modern times, however, the same word was written as 土圭, a compound of 土 /to/ 'earth' and 圭 /kei/ 'pyramid-shaped jade'. Historically, this older form was used for writing /tokei/, originally denoting Chinese terracotta sundials. After the introduction of Western mechanical clocks to Japan in the mid-16th century, it was gradually replaced by various other forms (e.g., 時計, 斗鷄, 斗影) to reflect the change in time measurement devices. The current 時計 became the only accepted form as a result of orthographic regularisation. It is difficult to see how to explain this orthographic change from 土圭 to 時計 from a purely morphological standpoint. The only possibility would be to assume two homophonous variants of the word /tokei/, consisting of different pairs of morphemes. The validity of such an assumption is open to discussion. For one thing, it is not immediately clear at what level of abstraction the word's referent (i.e., time measurement device) can be considered to have different meanings (i.e., 'sundial' versus 'clock'). For another, the change in orthographic form (i.e., 土圭 > 時計) and lexical meaning (i.e., 'sundial' > 'clock') does not necessarily entail a change in the word's morphemic make-up (i.e., {earth} + {pyramid-shaped jade} > {time} + {measure}).

4. The morphophonetic theory

An alternative view has been suggested by the morphophonetic theory, which characterises kanji graphs as being primarily phonographic and only secondarily morphemic. This section discusses the reasoning behind this claim.

4.1. An Overview of the Morphophonetic Theory

DeFrancis (1989, pp. 47–64, 89–121) provides perhaps the strongest criticism of the notion of morphography as a major type of writing. The author argues that the most fundamental principle underlying all full-fledged writing systems is phonography, which may or may not be sup-

plemented by a limited number of non-phonographic signs. For DeFrancis, this is also applicable to the Chinese writing system, which is traditionally considered as a prime example of logographic or morphographic writing systems. Matsunaga (1994, pp. 20–39; 1996, pp. 14–18) follows the same line of argument and characterises Japanese kanji graphs as being *morphophonetic*, that is, primarily phonographic and only secondarily morphographic (see footnote 4 above). For the purpose of the present paper, let us first take a closer look at DeFrancis' treatment of Chinese, and then proceed to examine Matsunaga's discussion of Japanese.

DeFrancis' (1989) argument for Chinese as an essentially phonographic writing system is based on two facts. The first one is that all graphs in Chinese, known as *hànzì* (traditionally 漢字 / simplified as 汉字), are associated with one or more monosyllabic readings, but not all readings convey constant meanings.¹⁷ Thus, while DeFrancis acknowledges that many readings indeed correspond to individual morphemes, he emphasises that *hànzì* graphs are primarily monosyllabic and only secondarily monomorphemic. The second—and more important—fact is that the majority of *hànzì* graphs are what DeFrancis terms *SP compounds*, that is, combinations of graphical components called *semantic* and *phonetic elements*. Roughly, the semantic element suggests the semantic class under which the graph's reading is traditionally classified, whereas the phonetic element indicates the way this reading should be pronounced. To take one of DeFrancis' examples, 像 is associated with the reading /xiàng/ 'image' in Chinese. This graph consists of the semantic element 亻, which derives from 人 /rén/ 'person,' and the phonetic element 象, which by itself represents the word /xiàng/ 'elephant'. Here, the former suggests the semantic class 'person' irrespectively of the reading /rén/, and the latter indicates the pronunciation /xiàng/ without regard to the meaning 'elephant'. According to DeFrancis, phonetic elements are found in about 97% of all Chinese graphs created by the 18th century. Taking these two facts together, DeFrancis argues that the Chinese writing system should be characterised as being *morphosyllabic*, that is, primarily syllabic and only secondarily morphographic. A similar view is shared by his predecessor Gelb (1963, pp. 85–89) and contemporaries like Unger (1987, pp. 35–49) and Daniels (1992, p. 83; 2018, pp. 84–92).

While DeFrancis (1989) stops short of clarifying whether the same characterisation is possible for Japanese kanji graphs, Matsunaga (1994; 1996) argues in favour of that position. Given the polysyllabic nature of kanji readings in Japanese, Matsunaga characterises kanji graphs as being *morphophonetic*, an umbrella term also proposed by DeFrancis to designate all writing systems that are primarily phonographic and secondar-

17. As an exception to the monosyllabic nature of *hànzì* graphs, 兒/儿 is read monosyllabically as /r/ when used to write the diminutive suffix *-r*.

ily morphographic. Matsunaga finds support for her argument in Itō's (1979, pp. 71–75) survey of 1,933 frequently used kanji graphs. The set of kanji graphs used in this survey included 1,850 graphs of the *Tōyō Kanji Hyō* (当用漢字表) or 'List of Characters for Current Use,' a predecessor to the current *Jōyō Kanji Hyō* (Section 2). According to Itō, the 1,933 graph set included 1,248 SP compounds, of which 1,192 graphs had clearly identifiable phonetic elements. Regarding this latter subset, she reports that 734 graphs (61.6%) had phonetic elements that would indicate pronunciations in an accurate way. Based on Itō's findings, Matsunaga maintains that the role of phonetic elements is as important in Japanese kanji graphs as they are in Chinese hànzi graphs.

4.2. Problems of the Morphophonetic Theory

Matsunaga's (1994; 1996) above argument provides important insights into the phonological aspect of the functioning of kanji graphs. Nonetheless, it places too much emphasis on the functionality of phonetic elements. There are two main problems with this.

Firstly, phonetic elements indicate only one of two types of readings. As will be recalled from Section 2, kanji graphs are typically associated with both *on* and *kun* readings, the former originating in Chinese and the latter in Japanese. Importantly, while phonetic elements indicate *on* readings more or less accurately in many SP compounds (see below), they do not provide any information about *kun* readings. For instance, 白 can both stand by itself as an independent graph as well as form a phonetic element in other graphs like 柏, 粕 and 泊. It provides an accurate indication of the *on* reading for these graphs, namely 白 /haku/ 'white,' 柏 /haku/ 'daimyo oak,' 粕 /haku/ 'dreg' and 泊 /haku/ 'stay'. With regard to *kun* readings, however, the same graphs are read in phonologically diverse forms, namely 白 /shiro/ 'white,' 柏 /kashiwa/ 'daimyo oak,' 粕 /kasu/ 'dreg' and 泊 /to(maru)/ 'stay'. As these examples clearly illustrate, phonetic elements may work for *on* readings but not for *kun* readings.¹⁸

Secondly, Itō's (1979) survey findings require careful re-evaluation. As already noted, Itō reports that 61.6% of the kanji graphs examined had phonetic elements that would accurately indicate pronunciations. It

18. There are some apparent exceptions in *kokuji* graphs (Section 2). For instance, 榎 is associated with the *kun* reading /masa(ki)/ 'Japanese spindle'. The graph incorporates 正, which can also stand by itself as an independent graph carrying the *kun* reading /masa/ 'exact' among other readings. Accordingly, this element may be considered as an example of phonetic elements indicating *kun* readings. However, in a discussion of 249 *kokuji* graphs, Sproat (2000, pp. 155–156) points out that only 8% of these graphs classify as SP compounds of this kind.

will also be recalled that this figure was obtained by dividing the number of SP compounds incorporating phonologically reliable phonetic elements (734 graphs) by the number of SP compounds with clearly discernible phonetic elements (1,192 graphs). Crucially, however, the percentage falls to 38.0% if one takes into account all the 1,933 kanji graphs used in the survey. This means that less than 2 in 5 graphs have a phonetic element accurately indicating *on* readings. In this light, the actual effectiveness of phonetic elements is also called into question even with regard to *on* readings.

In this connection, it is useful to consider two similar surveys conducted by later studies. The first one is Nomura and Itō's (1978, pp. 308–310) reanalysis of the 1,850 graphs of the *Tōyō Kanji Hyō*, which were part of the 1,933 graph set used in Itō's (1979) survey.¹⁹ According to Nomura and Itō, they included 1,137 SP compounds incorporating clearly discernible phonetic elements. However, when it comes to accuracy, these elements indicated the exact pronunciations of *on* readings in less than 1 in 3 graphs (33.3%). This result shows an even lower estimate for the effectiveness of phonetic elements than the one reported in Itō's earlier study. The second survey is presented in Stalph's (1989, pp. 148–155) study of kanji graphs and readings. Stalph points out a methodological problem in the two previous studies. In a nutshell, both Itō (1979) and Nomura and Itō (1978) identified phonetic elements and their corresponding pronunciations based on historical usage, and then compared them directly with their present-day counterparts. Criticising this confusion of synchrony and diachrony, Stalph presents a strictly synchronic analysis of 1,945 kanji graphs included in the pre-revision version of the *Jōyō Kanji Hyō* or the List of Characters for General Use (Section 2). The author reports that this set included 310 SP compounds (16.0%) containing a phonetic element indicating the exact pronunciation of an *on* reading. This figure casts further doubt on the notion that phonetic elements play a significant role in kanji graphs.

To summarise, Matsunaga (1994; 1996) is right in pointing out the prevalence of SP compounds and the existence of functional phonetic elements. However, the actual effectiveness of phonetic elements is virtually non-existent with respect to *kun* readings and highly limited in relation to *on* readings. In this light, it is implausible to characterise kanji graphs as being morphophonetic or primarily phonographic on the basis of phonetic elements. By doing so, one confuses the historical formation principle underlying kanji graphs and the way these graphs function in the current Japanese writing system.

19. The survey reported in Itō (1979) was conducted before the publication of Nomura and Itō (1978).

5. A New Proposal

Following the discussion presented in Section 3 and Section 4, it is now possible to establish the pros and cons of the existing theories. On one hand, the morphographic theory excels in capturing the fact that many kanji graphs correspond to individual morphemes. However, it is misleading to suggest that the morphographic principle underlies all kanji-written words except for those formed by phonetic borrowing. For one thing, it is dubious to assume morphological constituency in two-kanji compound words with a low degree of compositionality. For another, it is unclear how to deal with synchronic and diachronic orthographic variation in which different kanji graphs are used to write the same word. On the other hand, the morphophonetic theory sheds light on the phonological aspect of kanji graphs and kanji-written words without saying what is exceptional and what is not. At the same time, it assigns too much importance to the role of phonetic elements, whose effectiveness is highly limited in actuality. In short, while these two theories provide important insights into the relationship between kanji graphs and the morphological and phonological aspects of Japanese words, both make questionable assumptions to prioritise one aspect over the other.

For a better and more holistic understanding of the way kanji function as written signs, the present paper proposes to combine the advantages of the existing theories while avoiding their disadvantages. This proposal consists of two central claims. The first one is that kanji graphs relate to morphology by way of phonology. This is motivated by the observation that morphological constituency is justifiable in some kanji-written words (e.g., 国道 /kokudō/ ‘national road,’ 変化 /henka/ ‘change’) but not in those formed by phonetic borrowing (e.g., 葡萄 /budō/ ‘grape,’ 面倒 /mendō/ ‘care’) and those with a low degree of compositionality (e.g., 勉強 /benkyō/ ‘study,’ 挨拶 /aisatsu/ ‘greeting’). What this means is that kanji graphs may or may not correspond to individual morphemes, while they always correspond to certain portions of words’ phonological forms. Crucially, this is true regardless of whether or not the graphs incorporate synchronically effective phonetic elements. To capture these points, it is reasonable to generalise that all kanji graphs represent the phonological exponents of morphemes in both polymorphic and monomorphemic words (Figure 1).

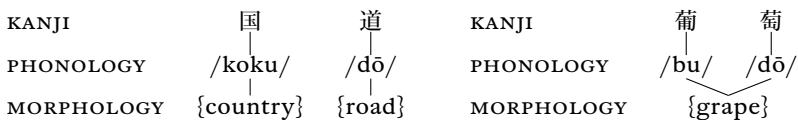


FIGURE 1. *Kanji* graphs representing phonological exponents of morphemes

The second claim is that kanji graphs function both individually and in fixed combinations. This is motivated by the existence of words written with single graphs (e.g., 書 /sho/ ‘writings’) and those written with multi-graph *jukuji* (e.g., 田舎 /inaka/ ‘countryside’). As both groups of words are generally monomorphemic (Honda, 2012, pp. 120–123, 128–133), it is fair to assume that kanji graphs can form two structurally distinct types of monomorphemic written signs, namely the single-graph and the multi-graph (Figure 2). Such a distinction is also justified by the prevalent use of polygraphs or multi-graph functional units across the world’s writing systems (Osterkamp and Schreiber, 2019).

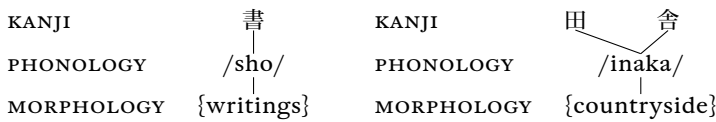


FIGURE 2. *Kanji* graphs forming single- and multi-graph written signs

Based on these claims, the present paper proposes a new, unifying model of kanji as written signs (Figure 3). In this model, kanji graphs are viewed as the formal building blocks of structurally simplex or complex written signs representing the phonological exponents of individual morphemes. The strength of the present model is that it provides a uniform account of the linguistic unit underlying the functioning of all kanji graphs without exception.

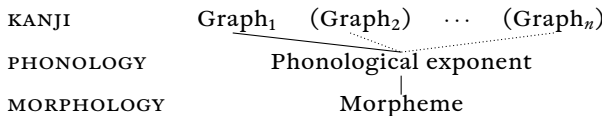


FIGURE 3. A unified model of kanji graphs as written signs

6. Concluding Remarks

This paper has discussed the type of linguistic unit represented by Japanese kanji graphs. After a preliminary discussion of the notion of linguistic unit (Section 1) and the relevant features of kanji graphs and kanji-written words (Section 2), it has presented a critical examination of the morphographic theory (Section 3) and morphophonetic theory (Section 4). Based on the discussion, the present paper has offered

a preliminary sketch of a new, unifying model of the way kanji function as written signs in the current Japanese writing system (Section 5). It has been proposed that kanji graphs can form structurally simplex and complex written signs, both representing the phonological exponents of individual morphemes. Further research is needed to test the validity of this proposal.

Acknowledgements

This study has been supported in part by the Centre for Language, Culture and Communication, Imperial College London. The author is deeply grateful to the programme committee and local organisers of the /gʁæfematik/ conference for giving me the opportunity to present an earlier version of this study. He also wishes to thank the participants to the conference for their valuable comments and suggestions. Special thanks are also due to the following people for their thoughtful feedback and constructive criticism: Prof. Jun Ikeda, Mr Edward Iles, Prof. Terry Joyce, Prof. Stefan Kaiser, Prof. Chieko Kanō, Prof. Hisashi Masuda, Prof. Akio Nasu, Dr Kazuhiro Okada, Prof. Sven Osterkamp, Dr Gordian Schreiber, Prof. Yoshiko Numata and Prof. Tadayuki Yuzawa.

References

- Backhouse, A. E. (1984). "Aspects of the Graphological Structure of Japanese". In: *Visible Language* 18, pp. 219–228.
- Bloomfield, Leonard. (1933). *Language*. London: George Allen and Unwin.
- Coulmas, Florian (1996). *The Blackwell Encyclopedia of Writing Systems*. Oxford, Malden: Blackwell Publishers.
- (2003). *Writing Systems: An Introduction to Their Linguistic Analysis*. New York: Cambridge University Press.
- Daniels, Peter T. (1992). "The Syllabic Origin of Writing and the Segmental Origin of the Alphabet". In: *The Linguistics of Literacy*. Ed. by Pamela Downing, Susan D. Lima, and Michael Noonan. Amsterdam, Philadelphia: John Benjamins, pp. 83–110.
- (1996). "The study of Writing Systems". In: *The World's Writing Systems*. Ed. by William Bright and Peter T. Daniels. Oxford: Oxford University Press, pp. 3–17.
- (2018). *An Exploration of Writing*. Sheffield: Equinox.
- Daniels, Peter T. and William Bright (1996). *The World's Writing Systems*. New York: Oxford University Press.
- DeFrancis, John (1989). *Visible Speech: The Diverse Oneness of Writing Systems*. Honolulu, HI: University of Hawai'i Press.

- DeFrancis, John and J. Marshall Unger (1994). “Rejoinder to Geoffrey Sampson, “Chinese Script and the Diversity of Writing Systems””. In: *Linguistics* 32, pp. 549–554.
- Frellesvig, Bjarke (2010). *A History of the Japanese Language*. Cambridge: Cambridge University Press.
- Gelb, Ignace J. (1963). *A Study of Writing*. Chicago: University of Chicago Press.
- Hill, Archibald A. (1967). “The Typology of Writing Systems”. In: *Papers in Linguistics in Honor of Leon Dostert*. Ed. by William M. Austen. The Hague, Paris: Mouton, pp. 92–99.
- Honda, Keisuke (2012). “The Relation of Orthographic Units to Linguistic Units in the Japanese Writing System: An Analysis of Kanji, Kana and Kanji-Okurigana Writing”. PhD thesis. University of Tsukuba.
- Itō, Kikuko [伊藤菊子] (1979). “形声文字と漢字指導 [SP Compounds and Kanji Education]”. In: *言語生活 [Language Life]* 326, pp. 68–81.
- Japanese Cabinet [内閣] (2010). 常用漢字表 [*A List of Characters for General Use*]. Japanese Cabinet Notification No. 2, issued on 30 November 2010.
- Joyce, Terry (2001). “The Japanese Mental Lexicon: The Lexical Retrieval and Representation of Two-Kanji Compound Words from a Morphological Perspective”. PhD thesis. University of Tsukuba.
- (2011). “The Significance of the Morphographic Principle for the Classification of Writing Systems”. In: *Written Language and Literacy* 14, pp. 58–81.
- Joyce, Terry, Hisashi Masuda, and Taeko Ogawa (2014). “Jōyō Kanji as Core Building Blocks of the Japanese Writing System: Some Observations from Database Construction”. In: *Written Language and Literacy* 17, pp. 173–194.
- Kaiser, Stefan [シュテファン・カイザー] (1995). “世界の文字・中国の文字・日本の文字: 漢字の位置付け再考 [Scripts of the World, China and Japan: Rethinking the Place of Kanji]”. In: *世界の日本語教育 [Japanese Language Education in the World]* 5, pp. 155–167.
- Kōno, Rokurō [河野六郎] (1994). “文字の本質 [The Essence of Writing]”. In: *文字論 [Theory of Writing]*. Ed. by Rokurō Kōno [河野六郎]. Tokyo: 三省堂 [Sanseidō], pp. 1–24.
- Kōno, Rokurō [河野六郎], Eiichi Chino [千野栄一], and Tatsuo Nishida [西田龍雄], eds. (2001). *世界文字辞典 [Encyclopedia of the World's Scripts]*. Tokyo: 三省堂 [Sanseidō].
- Kōno, Rokurō [河野六郎], Hidemasa Nagata [永田英正], and Hiroyuki Sasahara [笹原宏之] (2001). “漢字 [Kanji]”. In: *世界文字辞典 [Encyclopedia of the World's Scripts]*. Ed. by Rokurō Kōno [河野六郎], Eiichi Chino [千野栄一], and Tatsuo Nishida [西田龍雄]. Tokyo: 三省堂 [Sanseidō], pp. 256–281.
- Körtvélyessy, Livia, Pavol Štekauer, and Július Zimmermann (2015). “Word-Formation Strategies: Semantic Transparency vs. Formal

- Economy". In: *Semantics of Complex Words*. Ed. by Laurie Bauer, Livia Körtevelyessy, and Pavol Štekauer. Cham: Springer International Publishing, pp. 85–114.
- Lamb, Sydney M. (1966). *Outline of Stratificational Grammar*. Georgetown: Georgetown University Press.
- Matsunaga, Sachiko (1994). "The Linguistic and Psycholinguistic Nature of Kanji: Do Kanji Represent and Trigger Only Meanings?" PhD thesis. University of Hawai'i.
- (1996). "The Linguistic Nature of Kanji Reexamined: Do Kanji Represent Only Meanings". In: *Journal of the Association of Teachers of Japanese* 30, pp. 1–22.
- Miller, Roy Andrew (1967). *The Japanese Language*. Chicago, London: The University of Chicago Press.
- (1986). *Nibongo: In Defense of Japanese*. London: The Athlone Press.
- Miyajima, Tatsuo [宮島達夫] (1973). "無意味形態素 [Meaningless Morphemes]". In: *ことばの研究 [Study of Language]*. Ed. by Kokuritsu Kokugo Kenkyūjo [国立国語研究所]. Tokyo: 国立国語研究所 [Kokuritsu Kokugo Kenkyūjo], pp. 15–30.
- Morioka, Kenji [森岡健二] (2004). *日本語と漢字 [Japanese Language and Kanji]*. Tokyo: 明治書院 [Meiji Shoin].
- Nelson, Andrew N. (1997). *The New Nelson Japanese-English Character Dictionary*. Singapore: Tuttle Publishing.
- NKDDHI [日本国語大辞典第二版編集委員会], ed. (2000–2002). *日本国語大辞典 [Great Dictionary of the Japanese language]*. 2nd ed. Tokyo: 小学館 [Shōgakukan].
- Nomura, Masaaki [野村雅昭] (1988a). "二字漢語の構造 [The Structure of Two-Kanji Sino-Japanese Words]". In: *日本語学 [Japanese Linguistics]* 7, pp. 44–55.
- (1988b). "漢字の造語力 [Word Formation Productivity of Kanji]". In: *漢字講座 1 – 漢字とは –*. Ed. by Kiyoji Satō [佐藤喜代治]. Tokyo: 明治書院 [Meiji Shoin], pp. 193–217.
- (1999). "字音形態素考 [Considerations on Sino-Japanese Morphemes]". In: *国語と国文学 [Japanese Language and Literature]* 76, pp. 1–10.
- Nomura, Masaaki [野村雅昭] and Kikuko Itō [伊藤菊子] (1978). "漢字の表音度 [Phoneticity in Kanji]". In: *計量国語学 [Mathematical Linguistics]* 11, pp. 306–311.
- O'Grady, William and Videya P. de Guzman (1997). "Morphology: The Analysis of Word Structure". In: *Contemporary Linguistics: An Introduction*. Ed. by William O'Grady, Michael Dobrovolsky, and Francis Katamba. London, New York: Longman, pp. 132–180.
- Okimori, Takuya [沖森卓也] (2011). *日本の漢字 1600年の歴史 [The 1600-Year History of Kanji in Japan]*. Tokyo: ベレ出版 [Beret Shuppan].

- Osterkamp, Sven and Gordian Schreiber (2019). “<Th>e Ubi<qu>ity of Polygra<ph>y and Its Significance for <th>e Typology of <Wr>iti<ng> Systems”. Paper presented at *The Association for Written Language and Literacy’s 12th International Workshop, Diversity of Writing Systems: Embracing Multiple Perspectives*, 26th March 2019, University of Cambridge (Cambridge, United Kingdom).
- Peirce, Charles Sanders (1931–1958). *Collected Papers of Charles Sanders Peirce*. Ed. by Charles Hartshorne, Paul Weiss, and Arthur W. Burks. Cambridge: Harvard University Press.
- Sampson, Geoffrey (2015). *Writing Systems*. Sheffield, Bristol: Equinox Publishing.
- Satō, Kiyoji [佐藤喜代治], ed. (1987–1989). 漢字講座 [*Lectures in Kanji*]. Tokyo: 明治書院 [Meiji Shoin].
- Satō, Kiyoji [佐藤喜代治] et al., eds. (1996). 漢字百科大事典 [*Encyclopedia of Kanji*]. Tokyo: 明治書院 [Meiji Shoin].
- Saussure, Ferdinand de (1916). *Cours de linguistique générale*. Ed. by Charles Bally and Albert Sechehaye. Lausanne, Paris: Libraire Payot and C^{ie}.
- Seeley, Christopher (2000). *A History of Writing in Japan*. Honolulu: University of Hawai‘i Press.
- Smith, Janet S. (Shibamoto) (1996). “Japanese Writing”. In: *The World’s Writing Systems*. Ed. by Peter T. Daniels and William Bright. New York: Oxford University Press, pp. 209–217.
- Sproat, Richard (2000). *A Computational Theory of Writing Systems*. Cambridge: Cambridge University Press.
- Stalph, Jürgen (1989). *Grundlagen einer Grammatik der Sinojapanischen Schrift*. Wiesbaden: Harrasowitz.
- Suzuki, T. (1975). “On the Twofold Phonetic Realization of Basic Concepts: In Defense of Chinese Characters in Japanese”. In: *Language in Japanese society*. Ed. by F. C. C. Peng. Tokyo: University of Tokyo, pp. 175–192.
- Tajima, Masaru [田島優] (2006). “表語文字としての漢字 [Kanji as Logographs]”. In: 朝倉漢字講座 2 —漢字のはたらき— [*Asakura Lectures on Kanji 2: The Workings of Kanji*]. Ed. by Tomiyoshi Maeda [前田富祺] and Masaaki Nomura [野村雅昭]. Tokyo: 朝倉書店 [Asakura Shoten], pp. 1–16.
- Taylor, Insup and M. Martin Taylor (2014). *Writing and Literacy in Chinese, Korean and Japanese*. Amsterdam, Philadelphia: John Benjamins.
- Unger, J. Marshall (1987). *The Fifth Generation Fallacy: Why Japan Is Betting Its Future on Artificial Intelligence*. New York, Oxford: Oxford University Press.
- (1990). “The Very Idea: The Notion of Ideogram in China and Japan”. In: *Monumenta Nipponica* 45, pp. 391–411.
- Unger, J. Marshall and John DeFrancis (1995). “Logographic and Semasiographic Writing Systems: A Critique of Sampson’s Classification”. In: *Script and Literacy: Reading and Learning to Read Alphabets, Syllabaries*

- and Characters*. Ed. by Insup Taylor and David R. Olson. Dordrecht: Kluwer Academic Publishers, pp. 45–58.
- Vance, Timothy J. (2002). “The Exception That Proves the Rule: Ideography and Japanese Kun’yomi”. In: *Difficult Characters: Interdisciplinary Studies of Chinese and Japanese Writing*. Ed. by Mary S. Erbaugh. Columbus: National East Asian Language Resource Center, Ohio State University, pp. 177–193.
- Yokosawa, Kazuhiko and Michio Umeda (1988). “Processes in Human Kanji-Word Recognition”. In: *Proceedings of the 1988 IEEE International Conference on Systems, Man, and Cybernetics, August 8–12, 1988, Beijing & Shenyang, China*, pp. 377–380.

On the Nature of Unmotivated Components in Modern Chinese Characters


Tereza Slaměňíková

Abstract. From an etymological perspective, the graphics of Chinese characters are in general supposed to encode at least semantic, but primarily both semantic and phonetic information concerning the recorded linguistic unit. This attribute of the Chinese writing system is often pointed out, even when referring to the composition of the graphemes used in modern Chinese signary. A careful look, however, at the individual characters suggests that, in view of the current meaning or sound of the characters, the relationship between the graphic and linguistic structure might be partly or entirely missing. This means, in other words, that apart from semantically and phonetically motivated components, unmotivated constituents can be identified in the composition of modern Chinese characters as well. Although the phenomenon of unmotivated constituents has been discussed in a number of grammatological studies, it is often viewed as a peripheral issue. This paper argues that these units deserve much more attention than they have so far received. Based on a new model of the classification system for Chinese characters, it demonstrates that there are two different types of unmotivated constituents to be distinguished, and thus it provides deeper insight into the characteristic features of the modern Chinese writing system.

1. Introduction

Chinese characters represent the oldest, uninterruptedly used, writing system in the world. Over the course of its development, the graphic form of the characters underwent radical changes influencing the basic characteristic features of the writing system. The corruption of the

The preparation of this paper was made possible with the support of the Fond podporu vědecké činnosti at the Faculty of Arts, Palacký University in Olomouc (FPVČ2017/16 Svět v sinogramech).

Tereza Slaměňíková  0000-0001-6929-7568
Department of Asian Studies, Palacký University in Olomouc
Czech Republic
tereza.slamenikova@upol.cz

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1.
Fluxus Editions, Brest, 2019, p. 209–226. <https://doi.org/10.36824/2018-graf-slam>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

graphics of the characters and the changes in the Chinese phonological system as well as in the semantic content of the recorded linguistic units contributed to the disruption of the originally motivated relationship between the graphic and linguistic structures (cf. Schindelin in this volume). The currently used system therefore represents a certain mixture of graphemes with a different level of motivation. Despite this fact, it is not unusual for researchers to employ the system created almost 2,000 years ago when categorizing the modern Chinese characters.

The so-called six-category classification was described in the oldest known Chinese grammatological study 说文解字 *Shuō Wén Jiě Zì* [The Meaning Explanation of Primary Characters and the Structure Analysis of Secondary Characters]. Its author, the Han-dynasty scholar 许慎 Xǔ Shèn, conducted a thorough graphemic analysis of more than nine thousand characters in order to explain the relationship between their graphics and the recorded linguistic units. The definitions used also placed each character into one of the categories referring to the constructional method employed by their creating. As the title of the book indicates, there are two basic types of characters to be distinguished: (1) those with a simple graphic form represented by 象形 *xiàngxíng* pictograms (schematic depictions of the objects they represent) and 指事 *zhǐshì* symbols (characters expressing a certain idea through symbolic strokes), and (2) those with a compound graphic form represented by 会意 *huìyì* ideograms (combinations of two or more semantic components) and 形声 *xíngshēng* phonograms (combinations of one semantic and one phonetic components).¹ The last two of the six categories mentioned in the Postface of *Shuo Wen Jie Zi*, i.e., 假借 *jiǎjiè* loanwords and 转注 *zhuǎnzhu* variants, refer to the new usage of an existing character² and as such are not identifiable in the definitions of characters.

Considering its significant role in the Chinese grammatological tradition, it is not surprising that Xu Shen's legacy continues to shape the field up until the present. One should not, however, overlook the fact that the application of the traditional six categories to the modern Chinese characters suffers from certain limitations. The main reason for this is the fact that a comprehensive synchronic description of the graphic form composition of the characters is impossible without involving a new type of constructional units in the system, i.e., units that do not provide a link to the pronunciation or to the meaning of the recorded linguistic units. There are consequently two more basic types of characters to be distinguished: (a) with partly, and (b) with entirely unmotivated graphics.

1. Grammatologist are not in full agreement regarding the English equivalents of the six categories. This paper adopts terms used by Uher (2013, pp. 297–303) in the English translation of the Postface to *Shuo Wen Jie Zi*.

2. For details see Dong (1994, pp. 26–27).

The etymological explanation, by means of the six-category system, has represented an immutable classification paradigm for hundreds of years. Although researchers have often differed significantly in their views, not only on the theoretical concept of the categories as such, but especially on the classification of the individual characters, they never exceeded the established dogma.³ An important shift in approach occurred in the first half of the twentieth century⁴, when the palaeographer 唐兰 Táng Lán realized that the solution to the grammatical development crisis could be found in reducing the number of categories. His idea of adapting the traditional system inspired other grammatologists to propose their own modifications to the traditional model, particularly of importance being those developed by 陈梦家 Chén Mèngjiā, 裘锡圭 Qiú Xīguī or 王宁 Wáng Níng.⁵

According to another grammatologist 王凤阳 Wáng Fèngyáng (1989, p. 490), it was again Tang Lan who first used the term 记号字 *jìhào zì* unmotivated characters while referring to graphemes with corrupt graphics. It was Wang Fengyang himself, however, who in his extensive study highlighted the fact that the graphic composition of a significant number of Chinese characters no longer maintained the connection with their pronunciation or meaning. The already mentioned paleographer Qiu Xigui also paid significant attention to the issue of graphic form demotivation. His particular contribution lies in identifying the increase in the number of unmotivated characters and 半记号字 *bàn jìhào zì* partly unmotivated characters⁶ as one of the three major changes in the devel-

3. For details see Qiu (1988, p. 103).

4. According to Su (2007, pp. 2–3), the increasing interest in the characteristic features of the currently used writing system at the beginning of the twentieth century had its roots in two main factors. It was first influenced by the efforts of young Chinese intellectuals to reform traditional Chinese society, since the modernization of the written Chinese language was one of their main requirements. The educational system radically changed as a result: the classical texts were replaced with texts written in colloquial language. The use of the written language was no longer the privilege of a limited amount of state officials and therefore the Chinese writing system had to face an increasing number of its users. Secondly, in light of the technological progress, a need arose to make the Chinese writing system accessible to new products, such as printing, typing machines, telegraphs, etc.

5. Descriptions of these classification models can be found in Tang (2001, pp. 59–98), Chen (2006, pp. 24–94, 256–258 & 354), Qiu (1988, pp. 97–204) and N. Wang (2001, pp. 63–82).

6. The literal translation of the Chinese word 记号 *jìhào* is ‘mark, sign’. In light of the fact that these English terms are overloaded with various meanings used in different contexts and thus might be rather misleading, the indirect expression ‘unmotivated constituent,’ which reflects the nature of these units in the Chinese writing system, was used in this paper. The derived terms of the unmotivated characters and partly unmotivated characters refer to graphemes whose graphics demonstrate no or only a partial relationship to the represented linguistic unit.

opment of the constructional composition of the characters as well as in putting these characters into the context of the general development of the Chinese writing system.

Although Qiu Xigui, as well as, for example, Wang Ning, discussed the issue of unmotivated and partly unmotivated characters in their studies, keeping them outside their basic classification scheme. The protagonists of modern Chinese grammatology, a new autonomous discipline that crystallized in the 1980s, undertook such an innovative step. 周有光 Zhōu Yǒuguāng, who is considered the founder of this newest branch of Chinese grammatology⁷, claimed that its interest lies strictly in the currently used form of the Chinese writing system, the so-called modern Chinese characters. The question as to how the graphics of modern Chinese characters relate to their pronunciation or meaning naturally ranks among its main topics. In this context, the most recognized model of the new classification was introduced by by one of the leaders of modern Chinese grammatology 苏配成 Sū Péichéng.⁸ The so-called new six categories have opened up an alternative synchronic perspective on the etymologically oriented traditional classification system.

Over the years, Su Peicheng introduced three different versions of the new classification. In the first one, published in the oldest edition of his book 现代汉字学 *Xiàndài Hànzìxué*¹ [Modern Chinese Grammatology] (1994, pp. 72–80), he distinguished seven categories of characters. The first three categories include characters whose graphic form can still be viewed as fully motivated. Two of them were adopted from the traditional categorization, i.e., ideograms and phonograms. The third one contained characters whose graphics somewhat depicts their current meaning, that is the principle employed in Xu Shen's pictograms and symbols. The other four categories were newly implemented into the classification system, together with a new type component, i.e., an unmotivated constituent.⁹ Two of these four categories are partly mo-

7. For details see Su (2007, p. 3).

8. Su Peicheng himself provides examples of two other classifications introduced by Qian (2001) and Hao (1994). Classifications similar to Su Peicheng's model, however, differing in the organization of the categories in terms of the subordination of several categories under a superior group with similar qualities can be found, for example, in studies presented by R. Yang (2008) or G. Gao (2002). Interesting models were also introduced by H. Yang and Zhu (1996) and Pan (2003).

9. The word 'constituent' in the English translation was chosen based on two factors. First, Su Peicheng distinguishes three types of 字符 *zifú*, basic constructional units called components in this paper. Two of them are attributive compounds with the head 符 *fú*, i.e., 音符 *yīnfú* phonetic components and 意符 *yìfú* semantic components, and the third one 记号 *jìhào* was created with a different word formation principle. Second, the terminology used is somewhat inconsistent. Moreover, not all of the researchers use different terms when referring to the structural and constructional decomposition possibilities of the characters (for details see below). This is, for ex-

tivated, since they are combinations of an unmotivated constituent and a semantic or phonetic component. Finally, two categories with a completely unmotivated graphic form can be identified. One of them consists of simple unmotivated characters indivisible into smaller graphic units, while the other one involves complex unmotivated characters composed of two or more unmotivated constituents.

Unlike the first version, the two revised versions only employ six categories of characters.¹⁰ In the second edition of *Modern Chinese Grammarology* (2001, pp. 93–101), the category of simple characters depicting the meaning was excluded. It is listed again in the third edition (2015, pp. 102–111) and the six-category arrangement is achieved this time through a fusion of the simple and complex unmotivated characters into one category. In this manner, Su Peicheng once again implemented all the first-version categories into the classification. What is rather unfortunate is the fact that he does not provide any explanation that could help one understand the reasons for the repeated modification of the classification system.

The system of the new six categories is an unquestionable shift towards a synchronic approach to the relationship between the graphic and linguistic structure. It is apparent that, over the course of time, Su Peicheng gave this issue serious consideration in order to achieve a more efficient classification system. Regrettably, however, one cannot fail to notice that the definitions of the basic constructional units are somewhat shallow. Su Peicheng is quite specific concerning the requirements for phonetic components, however, the parameters that need to be present for a graphic unit to be considered a semantic component or an unmotivated constituent are too general to provide the required information for evaluation. It should also be mentioned that while describing the classification system, Su Peicheng avoids stating how many characters belong to each category. One might therefore ask whether he actually verified the applicability of the proposed system through an in-depth analysis of a representative sample of modern Chinese signary.

As the title indicates, this paper primarily focuses on the issue of unmotivated constituents.¹¹ The problem concerning their definition is connected with the fact that modern grammarology emphasizes the im-

ample, the case of Wang Ning who uses one universal term 构件 *gòujiàn* and thus does not establish so clear a line between the structural or constructional approach. The word ‘constituent’ is supposed to prevent over-interpretation in terms implying one or the other approach.

10. One cannot fail to notice that, through the exclusion of one category, Su Peicheng has reached an identical number of categories as can be found in Xu Shen’s classification system.

11. To describe in simple fashion the difficulty relating to semantic component, it is the rejection of the diachronic approach, one of the basic requirements of modern Chinese grammarology (Zhou 2004, pp. 306–316; Su 2001b, pp. 92–93; Su 2001a,

portance of a setting strict boundary between the constructional and the structural approach to the decomposition of Chinese characters. The first one explores the connection between the graphics of the characters and the meaning or sound of the recorded linguistic unit. The structural approach, in contrast, is strictly interested in character graphics. It examines the number, typology and arrangement of the minimal graphic units, i.e., strokes, and basic graphic units, i.e., (graphic) elements.¹² In view of this, the status of unmotivated constituents seems to be somewhat problematic since it is a constructional unit but carries no useful information concerning the character's meaning or pronunciation. Considering the graphemic analyses conducted by other researchers, the lack of a definition concerning the unmotivated constituent lies in the fact that it does not specify how deep the decomposition is supposed to be carried out and thus an uncontrolled blending with the structural decomposition methods is inevitable.¹³

This paper discusses the issue of unmotivated constituents on the basis of a new classification model that was proposed considering the above-mentioned limitations. The model was developed as an attempt to provide a system that will as effectively as possible reflect the current features of modern Chinese characters.¹⁴ It should be emphasized that it has been primarily associated with searching for a solution to the issue of unmotivated constituents. To put it into a time context,

pp. 359–360; Su 2015, pp. 101–102), that raises the question as to whether it is still an etymological explanation that should be used as an evaluation device.

12. Different terms are used when referring to the basic units of these two approaches: the element is the basic unit of the structural approach, while the component is the basic unit of the constructional approach. The term (graphic) unit is used as a general term for part of the character without any further implications.

13. In this respect, 快速识字字典 *Kuàisù Shí Zì Zìdiǎn* [Chinese Characters: Quick and Easy] (H. Yang and Zhu, 1996) can be taken as a representative example. As the title indicates, it is a dictionary and as such provides a graphemic analysis of modern Chinese graphemes, not “merely” proposing a certain kind of theoretical model, such as the above mentioned classifications mostly do. Although it demonstrates a high level of systematicity and undoubtedly serves its pedagogical purpose well, in case of the unmotivated parts, the authors tend to decompose them into minimal possible graphical units, such as, for example, the character 帶 *dài* ‘belt’ which is according to the authors composed of the semantic component 巾 ‘cloth’ and two unmotivated constituents 卅 (p. 47); or the character 令 *lìng* ‘command’ which is composed of three unmotivated constituents 人、 and 冫 (p. 164). The question that arises is whether it is, in the case of the constructional approach, reasonable to decompose characters into such small parts.

14. The new model was introduced in my dissertation at Palacký University in Olomouc, which has been published, in a revised version, under the title *Čínské znakové písmo: synchronní model tradiční kategorizace* [The Chinese Writing System: A Synchronic Model of the Traditional Categorization] (2017). A brief description of the model in English can be found in Slaměňíková (2017).

it was elaborated before the third version of Su Peicheng's categorization was published, i.e., at the moment when the simple and complex unmotivated characters each had its own separate category. The new model is based on the graphemic analysis of the so-called 2,500 常用字 *chángyòng zì* frequently used characters.¹⁵ During the analysis procedure, it was discovered that due to the extent and diversity of the corruption of the original form, one cannot avoid employing both the constructional and structural approach while categorizing modern Chinese characters. Only a combination of both approaches enables the establishment of a comprehensive classification. Nevertheless, when considering the primary interest in examining the relationship between the graphic and linguistic representation, the constructional principle is considered the superior one, while methods of structural decomposition are applied as supplementary tools to make an adequate processing of all the characters possible. As will be demonstrated, an important difference from Su Peicheng's categorization lies in the fact that the two approaches do not blend together, but either one or the other is applied.

2. Two Types of Unmotivated Constituents

The new model of categorization has a two-dimensional arrangement. It includes five groups subdivided into 20 categories. The group status reflects the decomposition specification, and the category status reflects the nature of the relationship between the entire character and its components in terms of semantic and phonetic motivation. It should be pointed that it was the unmotivated parts of the graphics of the characters that significantly determined its final arrangement. By means of the analysis, two types of unmotivated constituents were identified:

- (a) those that are not motivated in a particular character, however, they are used as phonetic and/or semantic components in other characters within the modern Chinese signary¹⁶. For example, the graphic

15. As concerns the representativeness of the sample, it should be pointed out that even though the analyzed characters cover less than one third of the currently used signary, represented by the 现代汉语通用字表 *Xiàndài Hànyǔ Tōngyòng Zìbiǎo* [Table of the Commonly Used Modern Chinese Characters] with a total amount of 7,000 graphemes, one cannot overlook the fact that, in the view of the high occurrence rate in modern Chinese texts, they undoubtedly stand for the core of the modern Chinese signary. From a qualitative point of view, the list of characters compiled on the basis of another classification criterion firstly does not eliminate the characters of a particular principle a priori, and secondly maintains a certain proportion of possible constructional principles.

16. To be specific, the occurrence within characters listed in *The Table of the Commonly Used Modern Chinese Characters* was taken into consideration.

- unit 巾 in the character 幫 *bāng* ‘help’ does not provide any useful link to the recorded morpheme, but is used as a semantic component in 帳 *zhàng* ‘curtain,’ 帽 *mào* ‘hat’ or 帆 *fān* ‘sail’. A certain situation can be observed in the case of the graphic unit 巨 in the character 柜 *guì* ‘cabinet’ which is, however, used as a phonetic component in the characters 距 *jù* ‘distance’ or 炬 *jù* ‘torch’
- (b) those that appear neither as meaning nor as pronunciation indicators at all, such as for example the graphic unit 𠄎 used in the characters 責 *zé* ‘duty,’ 素 *sù* ‘plain’ or 青 *qīng* ‘green’; or the graphic unit 兒 used in the character 貌 *mào* ‘appearance’.

In order to describe the two types of unmotivated constituents in detail, there is a need to pay attention to the motivated parts of the characters first. The synchronic point of view created a need to reconsider the definition of what kind of graphic units can be labeled as components. Apart from specifying the requirements on what is considered a semantically or phonetically motivated part of the character, one more parameter was added to the synchronic definition of the component, this being the recurrence. This means that only graphic units that occur as a semantic or phonetic indicator in at least two characters in the modern Chinese signary, represented by the 7,000 commonly used characters, were considered components.¹⁷

Various studies were consulted in order to define the criteria that a graphic unit had to meet to be considered an effective phonetic component in relation to the current pronunciation of the character or to be considered an effective semantic component in relation to its current meaning. As concerns the phonetic motivation, a great variability of methods, achieving noticeably different results, were observed. Simply speaking, the two basic approaches can be identified, when considering the graphic level which is being targeted. The first one focuses on phonetics that are classified according to the relationship between their syllabic value and the syllabic value of all the characters where they occur.¹⁸ The second approach, in contrast, functions the other way around. It focuses on characters since it examines whether a character contains a component indicating its pronunciation. These characters are there-

17. Including their occurrence on the higher constructional level, i.e., their occurrence as characters.

18. The following basic types of phonetics are distinguished: (a) ideal phonetics whose pronunciation is identical with all the characters they occur in; (b) phonetics with regular differences whose pronunciation deviate in a systematic manner; and in case of an indulgent approach also (c) irregular phonetics with an unsystematic relationship to the character’s pronunciation. This approach was employed, among others, by J. Gao, Fan, and Fei (1993); Zhang (1992); Guder-Manitius (1999); Schindelin (2007); Haralambous (2013).

fore sorted based on the level of phonetic component effectiveness.¹⁹ Unfortunately, what is common to both approaches is that researchers significantly differ in terms of the required level of syllabic value adequacy between the character and its phonetics. While some of them recognize as phonetics only those components that share exactly the same pronunciation as the character or differ no more than in tone, the other considers a correspondence either in the initial syllable or the final acceptable. Considering the target of this paper, the second approach was adopted when analyzing the phonetic motivation. As for the required syllabic value adequacy, graphic units with a correspondence at least in the initial or final were considered phonetically motivated. The reason for adopting this broader viewpoint was the fact that even in Xu Shen's Postface the required level of adequacy is not specified.

When comparing the phonetic motivation, the nature of the synchronic connection between the character graphics and the meaning does not appear to be examined almost at all. Although a wide spectrum of different handbooks can be found analyzing the semantic relationship between single characters and their components²⁰, the general theoretical implications are rarely discussed. This was the reason why a complex semantic characterization of each component was developed to evaluate the semantic link between a component and the meaning of a particular character.²¹ In order to achieve this, characters with the same component were gathered together and the meaning of each character²² was compared with the component's meaning as described in the grammatological dictionaries.²³ It has been observed that the same type of connection often repeatedly occurs in characters with the same component. To provide an example, under characters with the component 钅 (金) 'metal,' there can be found those referring to the following four main semantic classes: types for metals (e.g., 铜 *tóng* 'copper,' 锡 *xī* 'tin,' 铅 *qiān* 'lead,') different metal objects (e.g., 锤 *chuí* 'hammer,' 镰 *lián* 'sickle' 锁 *suǒ* 'lock,' 钉 *dīng* 'nail,' 锣 *luó* 'gong,' 锅 *guō* 'pot,' 链 *liàn* 'chain'), activi-

19. This approach was taken, among others, by Zhou (1980); Wen (1987); Defrancis (1984); H. Yang and Zhu (1996); Li and Kang (2002).

20. See e.g., Ye (2008); Huang and Ao (2009); H. Yu and Ch. (2010). Nevertheless, speaking of the semantic relatedness between characters and their constituents, an interesting approach of Haralambous (2013) has to be mentioned, who introduced an enhanced model for sinographic language processing. By exploring the semantic information stored in the so called subcharacters, he used three different WordNets.

21. This approach was inspired by the study of Shi (1992, pp. 76–92).

22. Specifically, the meanings mentioned in the two following dictionaries were considered: 现代汉语词典 (汉英双语) *Xiàndài Hànyǔ Cídiǎn (Han-Ying Shuangyu)* [The Contemporary Chinese Dictionary. Chinese-English Edition] 2002; and 新华字典 *Xīnhuá Zìdiǎn* [Xinhua Dictionary] 2011.

23. Two dictionaries in particular were used: *Hanzi Xing Yi Fenxi Zidian* (Cao and Su, 1999) and *Kuaisu Shizi Zidian* (H. Yang and Zhu, 1996).

ties connected with the use of a metal object (e.g., 锻 *duàn* ‘forge,’ 销 *xiāo* ‘melt,’ 铸 *zhù* ‘cast, found’) and qualities of metal (e.g., 锐 *ruì* ‘sharp’).

Thus, based on the component distribution, a set of repeatedly used connections was identified representing the core of the component’s semantic network. The semantic picture obtained in this manner was used to evaluate the motivation of the rest of the characters with this component. This method was applied in case of all characters containing one semantic component. When considering the complex nature of the semantic link in characters composed of two or more semantic components²⁴, the already mentioned grammatological dictionaries were used to determine their motivation. An emphasis, however, on a clear link with the current meaning of the characters was placed in the evaluation process.

Following the described procedure, a set of 613 phonetically used graphic units and a set of 270 semantically used graphic units were identified. Since some of them may possess both a semantic and phonetic function, the final list of all components includes 778 items. When considering their function in particular characters, four types of components can be distinguished: (a) phonetic components, abbreviated as the p-component; (b) semantic components, abbreviated as s-components; (c) components that provide both a semantic and phonetic link to the morpheme represented by the character, abbreviated as s/p-components²⁵; and finally (d) graphic units listed in the set of 778 components that in a particular case of occurrence do not possess any semantic or phonetic function, i.e., it can be argued that this function was neutralized, this being the reason for calling them n-components.²⁶

24. The complexity of the semantic link in these characters lies in the fact that it may not be derivable from the separate meanings of the individual components. This is because it is encoded as a combination of two semantic entities and thus associated with a more complex association process. For details, see Slaměniková (2013).

25. Components with both a semantic and phonetic function can be found in *Shuo Wen Jie Zi* (Xu, 1963) even though none of the definitions of the categories mentions how the characters containing this component should be evaluated. It is therefore not surprising that there is no consensus regarding their classification: some researchers consider them ideograms, others phonograms (Dong, 1994, p. 21). A previous analysis of ideograms has shown that in the case of 68% commonly used characters, which are according to *Hanzi Xing Yi Fenxi Zidian* composed of two semantic components, one of these is classified as a phonetic component as well (Slaměniková, 2013). Thus, considering the fact that the combination of one semantic component and one component with both a semantic and phonetic function is more common than the combination of two semantic components, it is my opinion that components with both functions should be considered as a specific type of components.

26. While providing examples of the characters of the categories described below, the four types of components are distinguished as follows: meaning can be found behind an s-component; pronunciation can be found behind a p-component; both mean-

N-components represent the above-mentioned first type of unmotivated constituents. Analysis has shown, however, that the principle of in/divisibility into components itself is not enough to create an effective classification system. A significant number of characters can be found with complex graphics, apparently composed of more than one graphic unit, yet, one of them or all of them, do not match the criteria to be considered a component. Another aspect had to therefore be implemented into the model in order to ensure that characters with an obvious different level of composition complexity will be divided into separate groups. This was the moment when the unmotivated units belonging to the second type were taken into consideration.

3. Unmotivated Constituents in the New Model

Based on the statistics provided by the Qing-dynasty scholar Wang Yun (G. Yu, 1995, pp. 51–58), 264 (2.8%) pictograms can be found, 129 (1.4%) symbols, 1,254 (13.4%) ideograms and 7,697 (82.3%) phonograms in *Shuo Wen Jie Zi*. It is apparent that most of the characters listed in Xu Shen's work were created as compositions of two components.²⁷ In view of this fact, all the characters were first examined in terms of the possibility of the decomposition into two components. Following the above described requirements on components, it has been determined that despite the graphic form corruption and other development changes, two-component arrangement still represented the dominant construction principle. The difference, however, is the wider range of possible combinations resulting from the four-type classification system of components. Altogether, seven different combinations were identified, each of them representing one category in the proposed classification model. The letter C refers to the fact that the two-component characters represent the third group in terms of the graphic form complexity (i.e., group C). The numbers in abbreviations indicate the absolute frequency of occurrence within the analyzed signary. To locate the position of the components in the character graphics, the following abbreviations are used: L for left, R for right, U for up, D for down, I for inside and O for outside.

ing and pronunciation can be found behind an s/p-component; no additional information can be found behind an n-component.

27. Based on Xu Shen's definitions, ideograms represent the only category of characters that can be composed of more than two components. The occurrence, however, of three or more-component ideograms is quite small. Specifically, an analysis of ideograms within the *Table of the Commonly Used Modern Chinese Characters* has shown that less than 5% of the 1,241 identified ideograms are composed of three or more components (ibid.).

Category C1: n-component + n-component (142 characters)

猜 *cāi* 'guess': L 犭 (犬) R 青

遗 *yí* 'inherit': I 贵 O 辶

Category C2: s-component + n-component (336 characters)

鹊 *què* 'magpie': L 昔 R 鸟 'bird'

海 *hǎi* 'sea': L 氵 (水) 'water' R 每

Category C3: p-component + n-component (166 characters)

辅 *fǔ* 'assist': L 车 R 甫 *fǔ*

常 *cháng* 'often': U 尚 *shàng* D 巾

Category C4: s/p-component + n-component (27 characters)

皇 *huáng* 'emperor': U 白 D 王 *wáng* 'king'

银 *yín* 'silver': L 钅 (金) *jīn* 'metal' R 艮

Category C5: s-component + s-component (188 characters)

库 *kù* 'warehouse'; O 广 'shed' I 车 'vehicle' (Cao and Su, 1999, p. 662)

岩 *yán* 'rock, cliff'; U 山 'mountain' D 石 'stone' (ibid., p. 295)

Category C6: s-component + p-component (939 characters)

筐 *kuāng* 'basket': U 竹 'bamboo' D 匡 *kuāng*

裙 *qún* 'skirt': L 衤 (衣) 'clothing' R 君 *jūn*

Category C7: s-component + s/p-component (139 characters)

箩 *luó* 'basket': U 竹 'bamboo' D 罗 *luó* 'net'—basket woven from grass (ibid., p. 610)

泡 *pào* 'bubble, blister': L 氵 (水) 'water' R 包 *bāo* 'bag'—bag full of water (ibid., p. 348)

The total value of 1,937 characters indicates that the two-component characters cover nearly fourth-fifths of the 2,500 frequently used characters. The most productive combinational principle is the connection between the p-component and the s-component. It can therefore be stated that the dominant constructional pattern of the minor script observed in Xu Shen's *Shuo Wen Jie Zi* is still being preserved in the writing system of modern Chinese. A significant decrease in occurrence cannot be overlooked, however, since the current percentage value is less than 40%. The second most productive category is represented by characters that combine one s-component and one n-component. This implies the significance of the unmotivated constituents in modern Chinese signary.

After sorting out the two-component characters, the rest of the characters were examined. Two more specific groups of characters were separated: group D including three or more-component characters²⁸ and group E including the so-called characters with zero meaning²⁹. Neither

28. For example the character 狱 *yù* 'lawsuit' which is composed of 讠 (言) 'words' between two 犭 (犬) 'dogs' expressing the meaning that two dogs are fighting each other (Cao and Su, 1999, p. 652).

29. Characters with zero meaning represent a specific group of graphemes that deviate from the general arrangement between graphic and linguistic units in Chinese

of the characters cover more than 1.5% of the analyzed signary. They are therefore only briefly mentioned in this paper. The remaining one fifth of the characters are going to be further discussed in the paper.

At first glance, these characters display a high diversity of composition complexity. Although the two-component decomposition based on the above described principles of motivation, is impossible, a number of them can be apparently divided into smaller graphic units that repeatedly occur in the Chinese writing system. To distinguish these characters from those with a single indivisible graphic form, the decomposition method of the structural approach was applied. Specifically, the dictionary 汉字信息字典 *Hànzì Xīnxi Zìdiǎn* [Dictionary of Chinese Character Information]³⁰ was used to determine the divisibility into graphic elements. It has been determined that almost exactly one half of these characters can be divided in two or more elements and one half cannot. The first mentioned are included in group B and the other in group A.

Within the analyzed sample, a total of 256 group B characters were identified. Despite the fact that the graphic units composing these characters do not meet the requirements that would enable the characters to be classified as a group C member, about one third of the group B characters contains one s-component, p-component or s/p-component. Four different categories can therefore be identified within the group B characters. The largest category is composed, however, of characters whose specification is only divisible into two or more elements. Considering the fact that the category status is supposed to reflect the nature of the relationship between the graphic and linguistic form, there was no need to separate the category of characters containing the n-component.

Category B1: divisible into two or more elements (158 characters)

能 *néng* ‘can, be able’: composed of elements 亠月匕匕

建 *jiàn* ‘construct’: composed of elements 廴聿

Category B2: divisible into two or more elements + contains one s-component (71 characters)

句 *jù* ‘sentence’: composed of elements 勹口, the second one functions as an s-component 口 ‘mouth’

since they, unlike other characters, do not carry any meaning. In order to do so, they need to be combined with another character and become part of a two- or more-syllable morpheme. They themselves are thus linked with the language only on the phonetic level. This is, for example, the case with the characters 菠 *bō* or 蜻 *qīng*. The first is used as part of two-syllable morphemes 菠萝 *bōluó* ‘pineapple’ and 菠菜 *bōcài* ‘spinach’; the second one is used along with another character with a zero meaning as part of the two-syllable morpheme 蜻蜓 *qīngtíng* ‘dragonfly’.

30. This dictionary was chosen since it represents the source from which modern Chinese grammatologist often quote statistical data about the structural composition of modern Chinese characters (e.g., Su 2001a, pp. 331–332, 350–352, 428; Su 2015, pp. 97–98; Ma 2013, pp. 85, 113–114, 220–221; R. Yang 2008, pp. 133–134).

骨 *gǔ* ‘bone’: composed of elements 冎月, the second one functions as an s-component 月 (肉) ‘flesh’

Category B3: divisible into two or more elements + contains one p-component (22 characters)

齿 *chǐ* ‘tooth’: composed of elements 止人口, the first one functions as a p-component 止 *zhǐ*

聚 *jù* ‘gather, get together’: composed of elements 耳又禾, the combination of the first and second element functions as a p-component 取 *qǔ*

Category B4: divisible into two or more elements + contain one s/p-component (5 characters)

眉 *méi* ‘eyebrow’: composed of elements 尸目, the second one functions as an s/p-component 目 *mù* ‘eye’

贵 *guì* ‘expensive’: composed of elements 贝虫, the first one functions as s/p-component 贝 *bèi* ‘shell’

The attribute connecting the A group characters is “indivisibility”. This is principally ensured through the one-element graphic structure. Although most of these characters originated as pictograms or symbols, understanding the current connection between their graphic form and the meaning of the recorded morpheme usually requires a more or less extensive etymological explanation. Only a small number of characters can be found about whose graphics it can be said that they distinctively reflect the recorded meaning (labeled as Category A2). These characters are characterized by constructional indivisibility which is superior to the structural decomposition possibilities. In addition, one more specific category can be identified: graphemes where another independently existing character with exactly one more or less distinctive stroke can be recognized (labeled as Category A3). In traditional classification, characters composed on this principle would be considered symbols, however, compared with these it is important to highlight the one-stroke difference between the initial and derived character. Although it was not the original intention, the graphemic analysis has shown that only in these cases can the initial character actually be recognized and as such provide significant information in relation to the meaning or pronunciation of the derived character.

Category A1 (215): one unmotivated element

马 *mǎ* ‘horse’

石 *shí* ‘stone’

Category A2 (30): pictographic or symbolic reflection of the meaning

田 *tián* ‘field’: Earth’s surface divided by water canals into small fields

一 *yī* ‘one’: one horizontal stroke; 二 *èr* ‘two’: two horizontal strokes;

三 *sān* ‘three’: three horizontal strokes

Category A3 (12): existing character ± one distinguishing stroke

- 本 *běn* 'root': horizontal stroke added to the lower part of the character
 木 *mù* 'tree' symbolizes the meaning 'root'.
 灭 *miè* 'extinguish': a horizontal stroke added on top symbolizes the
 object that covered 火 *huǒ* 'fire'

4. Conclusion

The new model of classification used in this paper was primarily designed as an attempt to find a systematic solution for the evaluation of unmotivated units occurring in the graphics of modern Chinese characters. This was achieved through the adoption of a two-dimensional arrangement that enables an evaluation of the decomposition possibilities of the characters and the relationship between the graphic and linguistic structure separately. The limitations of Su Peicheng's new six categories were resolved, however, through a better specification of the principles applied by an evaluation of the semantic and phonetic motivation of the graphics of the characters from a synchronic perspective. In addition, the new model argues that only the implementation of both the constructional and structural decomposition methods can establish a good set of criteria for an effective classification system. It is important, however, to notice that these approaches do not blend together, but either one or the other is applied at a certain stage of decomposition process.

Thanks to the precisely defined parameters, two types of unmotivated constituents were identified, considering the potential of being used as a semantically or phonetically motivated graphic unit in other characters of modern Chinese signary. There are consequently different types of characters with entirely, and with partly unmotivated, graphics which can be distinguished. As can be observed below, the first mentioned can be divided into three types and the second one into two types.

Types of unmotivated characters:

- (a) indivisible represented by category A1;
- (b) divisible into elements represented by category B1;
- (c) divisible into two n-components represented by category C1.

Types of partly unmotivated characters:

- (a) divisible into elements represented by categories B2, B3 and B4—an s-component can be identified in the composition of the B2 characters, a p-component in the composition of the B3 characters and an s/p-component in the composition of the B4;
- (b) divisible into two-components represented by categories C2, C3 and C4—together with one n-component, the C2 characters are composed of one s-component, C3 characters of one p-component, and C4 characters of one s/p-component.

As mentioned above, the model was proposed before the third revised edition of *Modern Chinese Grammarology* was published. It should be admitted that, compared to Su Peicheng's classification, the proposed model has a rather complicated arrangement. It is, however, a two-dimensional layout that provides deeper insight into the nature of the graphemes composition and thus provides a better understanding of the characteristic features of the modern Chinese writing system. In closing, it is important to note that the proposed model, in its current form, represents a prototype that aims to provide the foundation for examining a larger data sample. Due to the fact that it is based on an analysis of only one part of modern Chinese signary, it has been developed with the intention of outlining the most complex spectrum of possibilities. Although the general pattern appears to sufficiently reflect the characteristics of the currently used writing system, certain modifications can be expected, especially considering the fact that groups D and E contain only a limited number of characters.

References

- Cao, X. [曹先擢] and P. Su [苏培成], eds. (1999). 汉字形义分析字典 [*Analytic Dictionary of Chinese Character Graphics and Meanings*]. Beijing: 北京大学出版社 [Beijing Daxue Chubanshe].
- Chen, M. [陈梦家] (2006). 中国文字学 [*Chinese Grammarology*]. Beijing: 中华书局 [Zhonghua Shuju].
- Chinese Academy of Social Sciences, Institute of Linguistics, Dictionary Department [中国社会科学院语言研究所词典编辑室] (2002). 汉英双语现代汉语词典 [*The Contemporary Chinese Dictionary: Chinese-English Edition*]. Beijing: 外语教学与研究出版社 [Waiyu Jiaoxue yu Yanjiu Chubanshe].
- Defrancisc, J. (1984). *The Chinese Language. Fact and Fantasy*. Honolulu, HI: University of Hawaii Press.
- Dong, X. [董希谦] (1994). 《说文解字》一夕谈 [*One Evening Talking about "Shuo Wen Jie Zi"*]. Zhengzhou: 河南人民出版社 [Henan Renmin Chubanshe].
- Gao, G. [高更生] (2002). 现行汉字规范问题 [*The Issue of Modern Chinese Characters Standardization*]. Beijing: 商务印书馆 [Shangwu Yinshuguan].
- Gao, J. [高家莺], K. Fan [范可育], and J. Fei [费锦昌] (1993). 现代汉字学 [*Modern Grammarology*]. Beijing: 高等教育出版社 [Gaodeng Jiaoyu Chubanshe].
- Guder-Manitius, A. (1999). *Sinographemdidaktik. Aspekte einer systematischen Vermittlung der chinesischen Schrift im Unterricht Chinesisch als Fremdsprache. Mit einer Komponentenanalyse der häufigsten Schriftzeichen*. Heidelberg: Julius Gross Verlag.
- Hao, E. [郝恩美] (1994). “现代汉字教学法探讨 [Investigation of Teaching Strategies for Chinese Characters]”. In: 语言文字应用 [*Applied Linguistics*] 1994, pp. 83–87.

- Haralambous, Y. (2013). "New Perspectives in Sinographic Language Processing through the Use of Character Structure". In: *Computational Linguistics and Intelligent Text Processing, CICLing 2013*. Ed. by A. Gelbukh. Vol. 7816. Berlin, Heidelberg: Springer, pp. 201–217.
- Huang, W. [黄伟嘉] and Q. Ao [敖群] (2009). 汉字部首例解 [*Chinese Radicals in Examples*]. Beijing: 商务印书馆 [Shangwu Yinshuguan].
- Li, Y. [李燕] and J. Kang [康加深] (2002). "现代汉语形声字声符研究 [A Study of Phonograms in Modern Chinese]". In: 现代汉字学参考资料 [*A Collection of Papers on Modern Chinese Grammar*]. Ed. by P. Su [苏培成]. Beijing: 北京大学出版社 [Beijing Daxue Chubanshe], pp. 141–154.
- Ma, X. [马显彬] (2013). 现代汉字学 [*Modern Grammar*]. Guangzhou: 济南大学出版社 [Jinan Daxue Chubanshe].
- Pan, J. [潘钧] (2003). 现代汉字问题研究 [*Analysis of Selected Issues of Modern Chinese Characters*]. Kunming: 云南大学出版社 [Yunnan Daxue Chubanshe].
- Qian, N. [钱乃荣] (2001). 现代汉语 [*Modern Chinese*]. Nanjing: 江苏教育出版社 [Jiangsu Jiaoyu Chubanshe].
- Qiu, X. [裘锡圭] (1988). 文字学概要 [*Outline of Grammar*]. Beijing: 商务印书馆 [Shangwu Yinshuguan].
- Schindelin, Cornelia (in this volume). "The Li-Variation (隶变/隸變) *libiàn*. When the Ancient Chinese Writing Changed to Modern Chinese Script".
- (2007). *Zur Phonetizität chinesischer Schriftzeichen in der Didaktik des Chinesischen als Fremdsprache. Eine synchrone Phonetizitätsanalyse von 6.535 gebräuchlichen Schriftzeichen*. Munich: iudicium Verlag.
- Shi, Z. [施正宇] (1992). "现代形声字形符表义功能分析 [An Analysis of Determinatives Semantic Function in Modern Chinese Phonograms]". In: 语言文字应用 [*Applied Linguistics*] 1992, pp. 76–92.
- Slaměňíková, Tereza (2013). *Ideogramy v moderní čínštině [Ideograms in Modern Chinese]*. Olomouc: Univerzita Palackého.
- (2017a). *Čínské znakové písmo: Synchronní model tradiční kategorizace [Chinese Writing System: A Synchronic Model of the Traditional Categorization]*. Olomouc: Univerzita Palackého.
- (2017b). "Proposal for a New Classification System for Modern Chinese Characters". In: *Studia Orientalia Slovaca* 16, pp. 87–106.
- Su, P. [苏培成] (1994). 现代汉字学纲要 [*Outline of Modern Chinese Grammar*]. Beijing: 北京大学出版社 [Beijing Daxue Chubanshe].
- (2001a). 二十世纪的现代汉字研究 [*20th Century Research in Modern Chinese Characters*]. Taiyuan: 书海出版社 [Shuhai Chubanshe].
- (2001b). 现代汉字学纲要 [*Outline of Modern Chinese Grammar*]. Beijing: 北京大学出版社 [Beijing Daxue Chubanshe].
- (2007). "现代汉字学的学科建设 [Founding of the Discipline of Modern Chinese Grammar]". In: 语言文字应用 [*Applied Linguistics*] 2007, pp. 2–11.

- Su, P. [苏培成] (2015). 现代汉字学纲要 [*Outline of Modern Chinese Grammarology*]. Beijing: 北京大学出版社 [Beijing Daxue Chubanshe].
- Tang, L. [唐兰] (2001). 中国文字学 [*Chinese Grammarology*]. Shanghai: 上海教育出版社 [Shanghai Guji Chubanshe].
- Uher, D. (2013). *Hanská grammatologie* [*Han Grammarology*]. Olomouc: Univerzita Palackého.
- Wang, F. [王凤阳] (1989). 汉字学 [*Chinese Grammarology*]. Changchun: 吉林文史出版社 [Jilin Wenshi Chubanshe].
- Wang, N. [王宁] (2001). 汉字学概要 [*Outline of Chinese Grammarology*]. Beijing: 北京师范大学出版社 [Beijing Shifan Daxue Chubanshe].
- Wen, W. [文武] (1987). “关于汉字评价的几个基本问题 [Several Basic Issues on the Evaluation of Chinese Characters]”. In: 语文建设 [*Language Planning*] 1987, pp. 7–12.
- Xu, S. [许慎] (1963). 说文解字 [*The Meaning Explanation of Primary Characters and Structure Analysis of Secondary Characters*]. Beijing: 中华书局 [Zhonghua Shuju].
- Yang, H. [杨洪清] and X. Zhu [朱新兰] (1996). 快速识字字典 [*Chinese Characters: Quick and Easy*]. Nanjing: 江苏古籍出版社 [Jiangsu Guji Chubanshe].
- Yang, R. [杨润陆] (2008). 现代汉字学 [*Modern Grammarology*]. Beijing: 北京师范大学出版社 [Beijing Shifan Daxue Chubanshe].
- Ye, Ch. [叶昌元] (2008). 字理——汉字部件通解 [*Chinese Characters Motivation: General Explanation of Characters Components*]. Beijing: 东方出版社 [Dongfang Chubanshe].
- Yu, G. [余国庆] (1995). 说文学导论 [*Introduction to The Meaning Explanation of Primary Characters and Structure Analysis of Secondary Characters*]. Hefei: 安徽教育出版社 [Anhui Jiaoyu Chubanshe].
- Yu, H. [余辉] and Liu Ch. [刘诚] (2010). 说字理解汉字 [*Explanation of Chinese Characters by Principles of Motivation*]. Beijing: 气象出版社 [Qixiang Chubanshe].
- Zhang, J. [张静贤] (1992). 现代汉字教程 [*Tutorial of Modern Chinese Characters*]. Beijing: 现代出版社 [Xiandai Chubanshe].
- Zhou, Y. [周有光] (1980). 汉字声旁读音便查 [*Guide to Phonetics Pronunciation in Chinese Characters*]. Changchun: 吉林人民出版社 [Jilin Renmin Chubanshe].
- (2004). “现代汉字学发凡 [Introduction to Modern Chinese Grammarology]”. In: 周有光语言学论文集 [*Collection of Zhou Youguang's Linguistic Articles*]. Beijing: 商务印书馆 [Shangwu Yinshuguan].
- 新华字典 [*Xinhua Dictionary*] (2011). 11th ed. Beijing: 商务印书馆 [Shangwu Yinshuguan].

The Li-Variation (隶变/隸變) *libiàn*. When the Ancient Chinese Writing Changed to Modern Chinese Script

Cornelia Schindelin

Abstract. In textbooks of Chinese as a foreign language as well as in other introductions to the Chinese script, the reader is often shown examples of Chinese characters in their modern form along with various historical forms to demonstrate how these characters evolved towards their present shape. When Chinese script is introduced in this way, it remains quite unclear whether the inventory as a whole or the relationships between character components and complete characters underwent any significant changes. However, as is well known at least to Chinese specialists in the field, in the 1st century AD, when the scholar Xu Shen wrote the first semasiological character lexicon of Chinese, changes within the Chinese script were already well under way which did not only alter the graphical appearance of Chinese characters but would eventually change the relationships among characters and the components contained in them. These changes are described and categorized in the present paper which aims at making this historical phenomenon better known to Western specialists in the field of graphemics.

1. Preliminaries

The aim of this paper is to better acquaint Western specialists in the field of graphemics with a development that took place in the Chinese script roughly two thousand years ago. This development is relevant because it comprises the evolution of the ancient Chinese script into the modern script people write today in China as well as in other parts of the sinophone world.

For the sake of brevity, a few presuppositions need to be made. The author shall assume that her readers basically understand how the modern Chinese script works even though they may not be competent in

Cornelia Schindelin
FTSK, Johannes Gutenberg-Universität Mainz
An der Hochschule 2 (Postfach 1150)
76711 Gernersheim
Germany
E-mail: schinc@uni-mainz.de

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings*
Grapholinguistics and Its Applications (ISSN: 2534-5192), Vol. 1.
Fluxus Editions, Brest, 2019, p. 227-243. <https://doi.org/10.36824/2018-graf-schi>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

reading it. Therefore, I shall take for granted that no further proof is necessary to show that Chinese characters are *not* an ideographic script. The Chinese writing system is a system whose symbols are in a certain way connected to the language they were constructed to record, and this connection is in the majority of characters phonetically motivated. The late John DeFrancis suggested to call the Chinese script a “morphosyllabic writing system” (DeFrancis, 1984, p. 88) to reflect the fact that in texts most characters (i.e., tokens) represent one morpheme corresponding to one syllable when read out loud. In dictionaries, of course, one and the same character (i.e., type) may be listed as a representative of potentially various meanings and even various corresponding syllables. The Chinese scholar Qiu Xigui, a grandseigneur of Chinese graphemics, insisted that the label attached to the Chinese writing system reflect the fact that the vast majority of characters are made up of components which serve certain purposes. His suggestion is to call it a “semanto-phonetic script” (Qiu, 2000, p. 13–28), in Chinese: 意符音符文字 *yìfú-yīnfú wénzì*, cf. Qiu (1988, p. 10–18).

Most Chinese characters belong to the category of signific-phonetic compounds¹, that is to say, they contain a signific component which gives a (rough) hint at the (original?) “meaning” of the character, while the other component gives a more or less useful hint at its pronunciation and can therefore be addressed as the phonetic component or, in short, the phonetic. These components may themselves be complex, and they may be able to “act” as complete characters themselves. Then there are other compound characters consisting only of signific components or of signific and purely mnemonic components which in another paper in this volume are called “unmotivated constituents” (Slaměniková, in this volume). And there are also simple characters; these in turn may show up as constituents in compound characters and serve as significs or phonetics, or even as mnemonic or unmotivated components. The ability to function as one or the other is not evenly distributed within the component inventory. A sensibly principled and structured analysis of a modern inventory of nearly 7,000 generally employed (simplified) characters as used in the People’s Republic of China will yield around 500 components (component types) (Bohn, 1998, p. 10–14). Fu Yonghe analyzed a larger inventory of 11,834 characters containing current sim-

1. The four—out of six (六书 *liùshū*)—traditional categories generally accepted as having been productive when new characters were needed are: 1. Pictographic characters (象形 *xiàngxíng*), making up about 4 percent of Xu Shen’s inventory; 2. Simple indicative characters (指事 *zhǐshì*), about 1 percent of Xu Shen’s inventory; 3. Compound indicative characters (会意 *huìyì*), about 13 percent of Xu Shen’s inventory; 4. Characters made up of a signific and a phonetic component (形声 *xíngshēng*), called “semantic-phonetic” by DeFrancis and “signific-phonetic” here, about 82 percent of Xu Shen’s inventory. (DeFrancis, 1984, p. 84) Besides DeFrancis (*ibid.*) or its German translation of 2011, Woon (1987) or Feng (1994) may serve as introductions.

plified ones as well as numerous characters which had not undergone simplification in the 1950's and counted 648 different components (Fu, 1993, p. 117).

Although Chinese characters from any age are fascinating to behold, I have refrained from including illustrations of the different scripts in this paper in order not to let it grow too thick. To make up for this, I shall attempt to provide useful search terms which should enable the reader to find relevant photographs and illustrations in the vast vaults of the world-wide web.

2. The Chinese Script before Li-Variation

The change within the character system called here Li-variation has been dubbed “watershed” and “milestone” by Chinese scholars. In order to appreciate this characterization, it is necessary to look at the script that was in use before the Li-variation set in. Considering the number of characters affected, the length of time this process took as well as the complexity of the entire phenomenon, the following remarks can only be extremely sketchy.

In ancient times, that is from the late second millennium to the early first millennium BCE, Chinese diviners wrote on the plastrons (belly side) of tortoise shells and scapula (shoulder bones) of oxen for pyromantic divination.² Later, archives of such “oracle bones” were buried and subsequently forgotten. Paleographic and archaeological investigation started no earlier than 1898 or 1899 when for the first time after several millennia pieces of bone with characters on them came to the attention of Chinese scholars interested in the matter.³

The plastrons and scapula show a script of pictographic origin with various degrees of iconicity. While certain pictographic characters are—because of their iconicity—easier to decipher than others, especially for scholars familiar with the material and spiritual culture of the time, an especially interesting fact to note is that examples for all four main categories of Chinese characters can be found on them, including signific-phonetic compounds, even though the proportion of characters of this category among all those characters that have been successfully deciphered is lower than in later periods of history (cf. DeFrancis 1984,

2. Other materials like pottery, stone, jade, horn and so on were also used but less frequently, it seems, cf. Qiu (2000, p. 60).

3. To see examples, do a picture search for “oracle bone inscriptions”. At the time of writing, using the search terms suggested in this paper yielded useful results. For ancient character specimens pay attention to the rubbings among the search results. They are usually taken from ancient artifacts while works of calligraphy on paper are more recent.

p. 84). Phonetic loaning also appears to have been employed in this early period.

Bronze vessels dating from the time of about 1300 BCE to the early first millennium BCE with inscriptions on them were found in addition to oracle bones. They also show the writing of the time. Bronze vessels from later times have been discovered as well, but the characters on them look different already.

Characters to appear on a bronze vessel can be worked into the mold or engraved into the metal surface after casting. Thus, the artisans' procedure to get a written character on a bronze vessel is not quite the same as that of someone who engraves characters on a tortoise plastron or a bone with the help of a pointed tool.⁴

In both cases the material for writing determined the execution of the characters, at least to a certain extent: Casting molds allow for round lines more easily than bony material or cold metal does; round shapes or enclosures in a mold can easily be "filled" and the modulation of lines is also quite easy, while on bone or cold metal it would mean tediously taking more material away, which is why engraved circles and enclosures are usually not "filled" and lines not modulated much. A clay mold can be corrected but if something is etched off a piece of bone or cold metal, it cannot be replaced. Time pressure and ease of execution were not an issue when these solemn pieces were produced. The modern notions of stroke and stroke order had not yet appeared. There is great variety in compound characters as "allographs" for writing the same word or morpheme show diverse arrangements of component parts, variety of relative size of component parts, varying numbers of components and so forth. The orderliness of arrangement of the whole text also varies and does not seem to have been a requirement.

Around the middle of the first millennium BCE, a form of script appeared which is now commonly called "Large (or Great) Seal script". While it can be described as a descendant of both the script found on oracle bones and that used on early bronze vessels, it does display certain characteristics to set it apart: It is written in rows of quite even width, a lot of lines within the characters are rounded to different degrees and even complete circles can be found. Still, a lot of variety remains among allographic versions of compound characters especially concerning the number of components and their spatial arrangement. This script, which was the official script of its time, was still quite tedious to write, too. The development of various economic and socio-cultural factors—among them the fact that the Zhou kingdom was disintegrating and seven smaller kingdoms strove to take its place—exerted a lasting pressure on the Chinese script.⁵

4. Picture search: "Chinese bronze bronzes characters".

5. Picture search: "large seal script bronze vessels".

Among the seven states of the ensuing era, the Warring States period (475–221 BCE), the state of Qin seems to have been a comparatively conservative one. In this state, the Large Seal script was used relatively conscientiously while unearthed texts from the other six states show various degrees of simplification and disintegration of the writing system. As time went on, the state of Qin overwhelmed the other six states one after the other and extended its administrative control over their territories. Whenever such a victory was complete, Qin made sure that in the new territory only its script was employed. By 221 BCE, the state of Qin had successfully overthrown the other six states. The first emperor of the newly unified China aimed at unifying his realm in all relevant aspects, and unification of the script was one of the measures to achieve this, the others applying to track gauge, weights and measures, and coinage. Paleographers, who investigate increasing numbers of datable bamboo slips and silk textiles with writing on them, tell us that Qin's script policy appears to have been quite successful. However, the Large Seal script was still too unwieldy for the demands of a vast empire led with the aid of a well structured bureaucratic administration, and, in fact, archeological excavations have yielded text finds in which the characters show mixed degrees of simplification. A solution to the script problem of the time was offered by high officials who standardized and further simplified the existing Seal script, resulting in what has come to be known as "Small (or Lesser) Seal script". Textbooks intended not just to promulgate knowledge but also to serve as models showing what each character should look like were produced by three high-ranking scholar-officials, and copies of these books were distributed everywhere in the empire.⁶

However, even while these efforts were under way, another development had started and was already gaining momentum.

3. The Li-Variation

This development which goes by the Chinese name 隶变 (trad. 隸變) *libiàn*,⁷ literally "scribes' variation"⁸, actually started sometime in the

6. Of course, readers may also find pictures with the help of the search term combination "lesser small seal script inscriptions," but only tracking the changes between Large Seal and Small Seal character allographs will reveal the actual differences between their forms.

7. This paper owes a lot to Zhao (2009). Other important sources are Qiu (1988; 2000), F. Wang (1989), and He, Hu, and M. Zhang (1995). To maintain readability and since the intended audience is expected to consist of people who are not practiced readers of Chinese, I have refrained from naming sources very often.

8. 隶/隸 *lì*: (of a human being) subject, subordinate, underling, serf, hence: scribe, clerk; 变/變 *biàn*: change, transform(ation). Several renderings of 隶变/隸

Spring-and-Autumn period (770–476 BCE) which owes its name to the title of the annals of one of the seven states which have been preserved and become a classic text.

The evolution of the Chinese script from these beginnings to the “modern” Chinese script took over 600 years and spanned the Warring States period, the Qin era when China was unified, and the Han era up to the break-up of the empire at its end in 220 AD. While the exact beginning may be debatable—since apparently no-one started the process intentionally and we cannot be sure if any of the earliest examples of this script are among the already unearthed specimens—, its end is to be found towards the late years of the Han dynasty when the Li-script 隶书 *lìshū*, which is what the Li-variation resulted in, was gradually replaced by the “regular script” 楷书 *kǎishū*, its elegant successor, which in fact is still used today.⁹ This latter process, however, is beyond the scope of this paper.

The change that later came to be called Li-variation started when people began to employ a kind of quick handwriting for writing down things of lesser official status or for private purposes. Since the official Seal script was slow and tedious to write, they took shortcuts to achieve greater writing speed and economy. The materials commonly used at that time were brushes, ink, and slips of bamboo, a very common material then, or other pieces of wood. Texts on textiles, especially silk weaves, have also been unearthed in graves dating in large part from the Han era (202 BCE–220 AD).

In the course of several centuries, formally slightly different styles of this handwriting style developed which shared many characteristics.¹⁰

變 *liàn* into English may be considered: “scribe’s/scribes” or “clerk’s/clerk’s change/transformation,” “Li-change” or “Li-transformation”. Zhao Pingan, in an article that seems to be a self-translation into English, uses the word “clericalization” (Zhao, 2009, p. 170–196) which might appear peculiar to Western readers. I prefer the renderings “Li-variation” and “Li-shift,” the latter because the phenomenon can be likened to phonological shifts in the sound system of a language. However, to retain closer resemblance to the Chinese term, I shall stick to “Li-variation” here.

All character readings I provide in this article, whether they be Chinese proper names, terms or character examples, will be modern pronunciations notated using the modern transcription system *Hanyu Pinyin*.

9. Search terms: “kai shu regular script”. In the People’s Republic of China 2,236 characters were further simplified in the 1950s into their now current forms. While this reform was dramatic enough for individual characters, it did not effect a deep-going shift within the whole system as the Li-variation had done “naturally” before it. For an example of a modern text, look up a popular online encyclopedia and select the Chinese version of an entry.

10. In fact, the development of the “running script” 行書 *xíngshū* started from Li-script, and it started quite early. “Running script” came about when Li-script characters were written even more hastily which resulted in further simplification by connecting and blurring strokes, in many instances keeping the contours of the character

While the simplifications and shortcuts used at the beginning seem random, the resulting Li-script eventually stabilized graphically and structurally. From a purely calligraphic point of view it is characterized by the fact that the characters show the existence of strokes in the modern sense of the term, are a little wider than high, although they usually each take up a hypothetical rectangle of the same size, and by the characteristics of their strokes. In certain styles of Li-script the last stroke is more pronounced, that is, it is thicker and drawn out a little longer than the other strokes of each character. Although there are angles, they usually do not appear as sharp as in the later “regular script” 楷书 *kǎishū* which is appreciated for its elegance, making Li-script characters look clumsier.¹¹

However, the style of strokes and the relative proportions of character components are only surface phenomena. What really makes this development so interesting are the changes that happened within the character system.

Several processes of change can be identified. Some of these primarily concern formal aspects of the characters, while others primarily affected them structurally. This distinction is partly artificial but it helps to break down the information and make this complex development accessible to our understanding.

3.1. Formal Changes

There was more than one process that affected the shape of the characters. Together these processes reduced the iconicity of characters at the graphical level.¹² Furthermore, they led to the evolution of the modern notion of “stroke” (笔画/筆畫 *bǐhuà*). These processes were:

but not completely writing out the details of each component and so forth. For a picture search use “li running script”.

11. To appreciate the stylistic differences between Li-script and “regular script,” try first doing a picture search for “han dynasty li script” and then another one for “wei dynasty kai script,” possibly in a new tab or register card, then compare. There are also books available which show the formal development of characters. L. Li (1992) treats 500 characters, most of them simple ones deriving from pictographs, so the stylistic changes are visible but not the systematic ones discussed in the next section. H. Wang (1993) discusses and shows a large number of simple and complex characters grouped in seven topical chapters. In most cases, the author presents more than one version of the same character from various script styles respectively. Although these books were written for laypeople and language learners, they provide a good glimpse at the formal variety of characters through history.

12. In fact, in 2014 and 2015 proposals were made to include Small Seal script characters in Unicode. The tables included in the 2015 proposal provide an opportunity to view large numbers of characters in their Seal script form and their modern appearance next to one another. See X. Li et al. (2015, p. 6–753).

- Straightening and angularization: Lines which had been round or bent to a certain degree in the Seal scripts, like bow-shaped lines and semicircles, were straightened out. So were lines in complete circles which were first broken up into semicircles and then straightened. Consequently, changes of directions even within one stroke (or what would become a stroke according to the modern notion of the phenomenon) which had been “round corners” became distinctively angular.
- Reduction: Quite a few characters lost one or more strokes or entire components. (See more below.)
- Junction: Lines which had been distinct and separate before now became joined, that is, they evolved into one stroke, in many cases a complex stroke involving an angle.
- Disjunction: In other cases, what had been one stroke before in the Seal script was broken up into two or more strokes in the Li-script.
- Addition: In some cases strokes were newly added to characters, possibly to improve their aesthetic balance.
- Repositioning: In some characters and character components strokes changed their place or rotated. In some cases complete components were rotated.
- Rounding or bending: There are not only cases of straightening but also of rounding. This mostly happened to lines that formerly had been slanted and not completely straight. During the Li-variation, certain slanting or curved lines developed into angular strokes.
- Changes in length: Both lengthening and shortening can be observed to have happened. These changes are owed to the fact that writers strove for evenness and balance both of the individual character and the entire text.

3.2. Structural Changes on the Level of Components

The following processes primarily affected the structure of compound characters and were not purely graphical. The addition of a stroke to a component for aesthetic reasons may result in this component changing its identity, such that one could also say that the former component was substituted with another one. However, it is not possible here—and not intended—to formulate and discuss criteria which could serve to separate cases of one kind from the other. We shall have to stay on a rather macroscopic and abstract level.

- Stabilization of the position of certain components, possibly with consequences at the graphical level: The graphical process of repositioning was already mentioned above. Repositioning is even more significant on the level of components. In the Seal scripts, allographs for the same grapheme (in the sense of whole character for a certain

word or morpheme) can be found which show that certain components could be written in various positions relative to one another without making a difference in meaning or pronunciation. In other words, the position a certain component could take up in “one and the same” character was not stable. Still, the component in question would have the same size and graphical shape in all its possible positions. This situation changed during Li-variation: Components that had formerly behaved unstable increasingly found a fixed position within the character or several characters they were constituents of, respectively. However, in many cases the same component ended up taking one position in one character and another in a different character. For many components, this process did not effect any significant changes on their shape, although some shift in relative size may have occurred; for others, the result was the development of allo-graphic components. These were not freely interchangeable, so eventually several different components resulted. The “heart” component 心 is a case in point: What had been one component before ended up as at least three: 心 (as in 想), 忄 (as in 情), and the four-stroke bottom component of 恭. As a result, readers and writers of Chinese must learn three shapes instead of just one for “heart”.

- Characters of the “signific + phonetic” category underwent still more changes on this level which also concerned the ability of their components to function as a signific or phonetic component.
 - Reduction of signific components: In the Seal scripts there were many characters whose signific component consisted of more than one minimal grapheme. During Li-variation, many of these lost some or all of the minimal graphemes making up the signific. In many cases, this made sense, especially where redundant components were eliminated. If at least one signific component was left, the resulting Li-character would still belong to the “signific + phonetic” category; otherwise it would then belong to another category or end up as one of those characters which are hard to categorize in the traditional system. This process could also happen to “signific + signific” characters of the compound indicative category.
 - Reduction of phonetic components: Several situations are possible. (1) If a part of the phonetic component was eliminated during Li-variation and the remainder gave no phonetic hint any longer, the resulting Li-character would no longer belong to the “signific + phonetic” category. It possibly became hard to categorize. (2) If the phonetic component itself was a character of the “signific + phonetic” category and a part of it was lost, the resulting Li-character could still belong to the “signific + phonetic” category if the remains of the component were able to function as a phonetic because it had been the phonetic part of the embedded

signific-phonetic character from the start. (3) If the phonetic component was simplified or reduced in the same way within all the characters it was a constituent of, taking up an identical shape in the resulting characters concerned, the resulting characters consequently would still belong to the “signific + phonetic” category and the new subcomponent would still function as phonetic.

- Substitution of the signific component: In certain characters, significs were substituted to achieve greater semantic transparency or writing economy. Researchers in China have identified groups of allographs with different significs that show that in the centuries of Li-variation it was by no means clear which of several eligible significs would be the best for certain characters, even if at the end one signific in each group may have gotten universally adopted. Some of the substitutions found in texts of the era in question result from confusion of graphically similar components. Others appear to be attempts to find the component that would best support the semantic transparency of the character.
- Substitution of the phonetic component: These substitutions probably happened to improve the fit between the reading of characters of the “signific + phonetic” category to contemporary pronunciation. This resulted in new series of characters sharing the same phonetic component.
- Addition of signific components: In many such cases the basis was a character of the simple or compound indicative category or the signific-phonetic compound category. The aim can usually be identified to be the creation of a character for a meaning (se-meme) formerly covered by the base character which had either been a phonetic loan or generally polysemous. Research into the Chinese character inventory and lexicon has shown that during the Han period the need intensified to write down words for which no characters had yet been developed. For a while the gap had been filled through extensive borrowing. However, later many of the phonetic loan characters were equipped with signific components which resulted in a considerable growth of the “signific + phonetic” category. This category was to remain the most productive one of the four.
- Complication of the phonetic component: In some cases the phonetic component became more complex by being exchanged for a complex character which contained the original component as one of its constituents.
- Exchange of a pictographic signific component for a phonetic component: Some of the resulting characters can be seen as consisting of two phonetic components, thus as having one component which serves both as phonetic and signific.

- Exchange of a pictographic component for a signfic one: As the iconicity of many characters decreased in the process of Li-variation, the loss of semantic transparency was at least partly compensated for by using established signfics instead of holding on to strokes with a formerly pictographic function from a time when the characters had been closer to pictography.
- Convergence of various combinations of components—possibly with varying functions in the respective original—to form a single new one. For the result, there are two possibilities: (1) The resulting component could function neither as a signfic nor as a phonetic; (2) It was able to function as a phonetic or signfic component.

3.3. A Look at One Group of Characters for Exemplification

To get an idea of the impact of Li-variation let us just look at one group of characters that were affected. What these characters have in common now is their top component. Before Li-variation their top halves had been composed of different component combinations, some of which had displayed a certain graphical similarity.

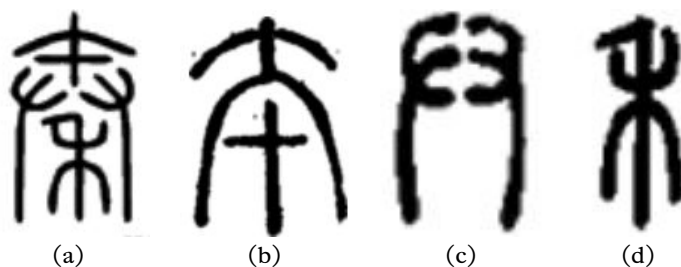


FIGURE 1. (a) 秦 (*qín*), form. (b) Top component (*tāo*), full Seal form. (c) Two hands with fingers pointing to the middle, Seal form. (d) Stalk of grain, Seal form

秦 (*qín*), Name of the state that unified China at the end of the Warring States period, 3rd century BCE (Fig. 1a): The top component of the Seal script version (Fig. 1b) is thought to have represented a pestle for grinding grain, beneath it there were two hands with the fingertips directed to the middle (Fig. 1c), the arms curving down to the left and right corner, respectively, and between the arms there is the character for “stalk of grain” (Fig. 1d). After the era of oracle bone inscriptions this character seems not to have been used for its original meaning (grain or

millet ready for grinding?), but only for the name of the empire of Qin and related names. A phonetic component cannot be identified.

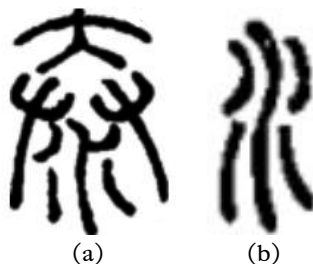


FIGURE 2. (a) 泰 (*tài*), Seal form. (b) 水 (*shuǐ*, water), Seal form

泰 (*tài*), peaceful, safe, very positive (Fig. 2a): The top component was the character 大 (*dà*, big, great), under its spread legs there were two hands with the fingertips directed to the middle (Fig. 1c), the arms curving down left and right, and between the arms there was the character for “water” in its ancient form (Fig. 2b). Here the top component served as the phonetic.

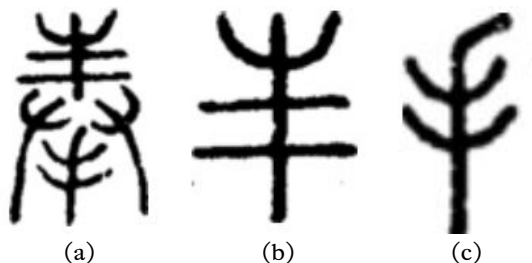


FIGURE 3. (a) 奉 (*fèng*), Seal form. (b) 丰 (*fēng*), Seal form. (c) 手 (*shǒu*, hand), Seal form

奉 (*fèng*), to present with both hands plus various meanings involving some kind of providing in a respectful way (Fig. 3a): Bronze inscriptions contain a simpler form of this character in which two hands offer a bundle of grain stalks (top part) representing abundance. The old top part was a character meaning “abundant”: 丰 (*fēng*) (Fig. 3b); it is identified as the phonetic component in this character. The Seal script shows a third hand (Fig. 3c) between the “arms” of the two hands, possibly to fill the space there and to reinforce the meaning. In the modern version of the character, this “hand” can be argued to occur in reduced form.

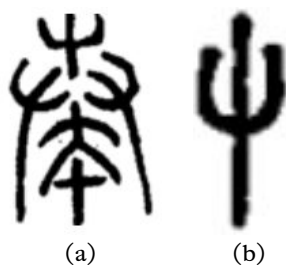


FIGURE 4. (a) 奏 (*zòu*), Seal form. (b) Top component 𠂔 (*chè*, plant sprout), Seal form

奏 (*zòu*), to perform, to effect (Fig. 4a): At the top there was a single plant sprout, i.e., a “rounder” version of 𠂔 (*chè*) (Fig. 4b), two hands underneath and the character (*tāo*) (Fig. 1b), to go forward quickly, which has fallen into disuse in the meantime, at the bottom between the “arms”.¹³

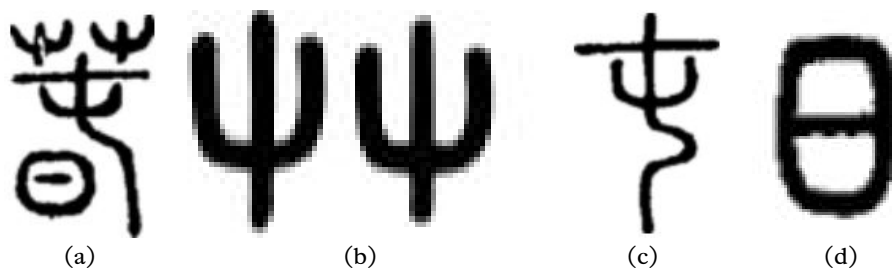


FIGURE 5. (a) 春 (*chūn*), Seal form. (b) The “grass” component, Seal form. (c) 屯 (*tún*), Seal form. (d) The “sun,” Seal form

春 (*chūn*), spring (Fig. 5a): This character in its Seal form comprised a component indicating “grass” at the top (a “rounder” version of the modern component, cf. Fig. 5b), the phonetic 屯 (*tún*)¹⁴ (Fig. 5c) in the middle, the last stroke of which curved down to the right, and to the left of this curving stroke a “rounder” version of the character 日 held to be a pictogram of the sun (Fig. 5d). This was a character of the significant-phonetic category. Its former phonetic component 屯 (*tún*) (Fig. 5c) is

13. It is interesting to note that the *Shuōwén* does not explain the component between the arms to be the phonetic of this character.

14. The *Shuōwén* explains this to represent a tender plant sprout having difficulties to push through the earth, thus meaning “difficult”. Thus, it may be argued that besides being the phonetic of the character, this component also supports its meaning.

still existent in the modern inventory and hints at pronunciations like *chun*, *dun*, and *tun*. Especially the components “grass” and “sun” may have helped to associate this character with its meaning of “spring time”.

This character is a bit particular, because after Li-variation one of its forms was a character of the same composition, just showing the graphical characteristics of the Li-script. However, as time went on and the “regular script” developed, the form 春 became popular and eventually superseded the Li-script form; a number of variants of this character also appeared, but 春 eventually was the form which won out as the orthographically accepted one. So in this case the fact that this character shares its top part with those discussed above cannot be solely attributed to the processes of Li-variation as evidently a certain degree of variation also went on when the Chinese writing developed into “regular script”.

As is discernible by comparing the Seal characters with the modern post-Li-variation versions of these characters¹⁵, the curved lines of the “arms” were straightened out and shortened, although they retained a certain slant to the left and right; the “fingers” were straightened and joined which resulted in horizontal strokes intersected by the left arm and the right arm joined beneath the third horizontal stroke; additionally, whatever had been atop the “hands” was melted together to form a horizontal stroke also intersected by the left-slanting “arm” stroke; and the strokes of the lower part were also straightened and angularized and—like in the case of 奉 (*fèng*)—simplified. The resulting bottom “hand” in the modern character 奉 (*fèng*) has a different form from the more common “hand” components 手 (*shǒu*, hand) and 扌 (the “upright hand” radical), while in the modern character 奏 (*zòu*) the bottom component now is 天 (*tiān*, heaven, sky; day) with a slightly varied right slanting last stroke due to the position it is placed in. The “grain stalk” in the modern character 秦 (*qín*) and the bottom “water” component in the modern character 泰 (*tài*), however, have assumed—or retained—the same forms as their counterparts elsewhere in the inventory.

So now, after Li-variation, and in the case of 春 (*chūn*) finally after the evolution towards “regular script,” these characters of different origins have a common top component. Three had the “two hands with fingertips directed towards each other” in common as well as the fact that there was something above the hands and something between the “arms” which used to extend to the bottom corners. Two of these three had had a phonetic component at the top which is not discernible any more today. The last character, 春 (*chūn*), in the end developed the same top part as the other ones, probably because the “grass” component (5b) graphically

15. There is another character with the same component at the top, 舂 (*chōng*, to grind something in a mortar), but discussing it would not add anything new to the argument.

somewhat resembled the two hands (1c) and the whole arrangement of components resulted in a similar outline of the character even though it originally lacked something like a “left arm”. This character lost its phonetic component when its “regular script” version developed.

All these characters in their modern forms must be memorized separately because their components do not tell the story of their (basic, original?) “meanings” nor give hints as to their pronunciation.

4. Summary

Especially during Qin and Han times (3rd century BCE through 3rd century AD), due to socio-cultural and economic reasons, the Chinese script underwent a profound change which led from the “old script” (古文 *gǔwén*) to the “modern script” (今文 *jīnwén*). Graphical changes occurred which among other things led to a loss of iconicity. In other cases, pictographically motivated traits were exchanged for components of established signfic function. In many characters, components were deleted, reduced, or substituted. Certain components lost their positional flexibility and assumed fixed positions within the characters they were constituents of. Certain (old) components split up into more than one new form, in effect becoming different (new) components. In other cases, various combinations of components melted together to form one identical new component devoid of the iconicity of its various forebears and not necessarily useful as phonetic or signfic component. A lot of new characters appeared which had no attested forerunners in Seal scripts or older inscriptions.

By the end of the Han period, the Chinese character system appears much more clearly than before as a system employing phonetic and signfic components of little iconicity, functional mainly by their association with certain pronunciations or “meanings,” respectively, to form characters of the “signfic + phonetic” category as the main units of its inventory. In fact, these characters comprise about 80 percent of the inventory at least since the first century AD (cf. DeFrancis 1984, p. 84).

When the resulting system was handed on and received by younger generations who were no longer familiar with the old Seal characters, the relationships between components were all the more perceived as they now appeared to hold. Thus, etymology with reference to the analyses in Xu Shen’s lexicon *Shuōwén-jìezì* (说文解字, Explanation of simple characters and analysis of complex characters; c. 100 AD) became an area of knowledge for specialists. Not everything about the Chinese writing system changed in the course of Li-variation: There is still a one-to-one relationship between morpheme, syllable and character in writ-

ten speech.¹⁶ People may disagree on the question whether “watershed” or “turning point” are adequate metaphoric expressions to characterize the Li-variation. However, even those who do not like these metaphors¹⁷ do not doubt that the Li-variation led to the development of the modern Chinese script.

Acknowledgements

I would like to thank my Chinese mentor and friend Prof. Wan Yexin 万业馨, Beijing, for her teaching, her critical comments and constant support during the past twenty-plus years. Not only have I greatly profited from our discussions about Chinese scripts, the character system, calligraphy and so forth, she also always finds a way to get a book or other material for me and to me when I need it for the next step in my research.

References

- Bohn, Hartmut (1998). *Quantitative Untersuchungen der modernen chinesischen Sprache und Schrift*. Hamburg: Kovač.
- DeFrancis, John (1984). *The Chinese Language: Fact and Fantasy*. Honolulu: University of Hawai'i Press.
- (2011). *Die chinesische Sprache: Fakten und Mythen*. Trans. by Stephan Puhl. Nettetal: Steyler Verlag.
- Feng, Zhiwei (1994). *Die chinesischen Schriftzeichen in Vergangenheit und Gegenwart*. Trans. by Wolfgang Kühlwein. Trier: Wissenschaftlicher Verlag.
- Fu, Yonghe [傅永和] (1993). “汉字结构和结构成分的基础研究 [Basic Research on the Structure of Chinese Characters and Their Structural Components]”. In: 现代汉用字信息分析 [*Analyses of Data about the Use of Modern Chinese Characters*]. Ed. by Yuan Chen [陈原]. Shanghai: 上海教育出版社 [Shanghai Jiaoyu Chubanshe], pp. 108–169.

16. Phoneticians would find reason to object to this—only slightly rough—characterization as modern Chinese speech provides cases where the pronunciation of certain two-character expressions results in one syllable phonetically. However, when native speakers read out text carefully, they usually read out one syllable for one character. Morphologists would also slightly object as there is indeed a small number of characters which stand for submorphemic syllables. In other words, the Chinese lexicon does contain some morphemes and simple words which need more than one character to write and more than one syllable to read out. The existence of these words—some of which are of venerable ancientness—let the fact that in the vast majority of cases a 1:1 relationship holds appear even more salient.

17. In fact, Zhao Pingan himself, on whom I rely heavily for this paper, doesn't like these metaphors, part of the reason being that he takes them too literally. (Cf. Zhao 2009, p. 71–72)

- He, Jiuying [何九盈], Shuangbao Hu [胡双宝], and Meng Zhang [张蒙] (1995). 中国汉字文化大观 [Grand Overview over the Culture of Chinese Characters]. Beijing: 北京大学出版社 [Beijing Daxue Chubanshe].
- Li, Leyi [李乐毅] (1992). 汉字演变五百例-修订版 *Tracing the Roots of Chinese Characters—500 Cases*. Beijing: 北京语言大学出版社 [Beijing Yuyan Xueyuan Chubanshe]. German version: *Entwicklung der chinesischen Schrift am Beispiel von 500 Schriftzeichen*, Beijing: Verlag der Hochschule für Sprache und Kultur, 1993; French version: *Évolution de l'écriture chinoise*, Beijing: Éditions de l'université des langues et cultures de Beijing, 1993.
- Li, Xian [李鑫] et al. (2015). "Proposal to Encode Small Seal Script in UCS". ISO/IEC JTC1/SC2/WG2 N4688 L2/15-281, <https://www.unicode.org/L2/L2015/15281-n4688-small-seal.pdf>.
- Qiu, Xigui [裘锡圭] (1988). 文字学概要 [An Outline of (Chinese) Grammatology]. Beijing: 商务印书馆 [Shangwu Yinshuguan].
- (2000). *Chinese writing*. Trans. by Gilbert L. Mattos and Jerry Norman. Berkeley, CA: Society for the Study of Early China.
- Slaměńíková, Tereza (in this volume). "On the Nature of Unmotivated Components in Modern Chinese Characters".
- Wang, Fengyang [王凤阳] (1989). 汉字学 [Grammatology of Chinese characters]. Changchun: 吉林文史出版社 [Jilin Wenshi Chubanshe].
- Wang, Hongyuan [王宏源] (1993). 汉字字源入门 *The Origins of Chinese Characters*. Beijing: Sinolingua. German version: *Vom Ursprung der chinesischen Schrift*, Beijing: Sinolingua, 1997.
- Woon, Wee Lee (1987). *Chinese Writing: Its Origin and Evolution*. Macau: University of East Asia.
- Zhang, Jingxian [张静贤] (1992). 现代汉字教程 [Course on Modern Chinese Characters]. Beijing: 现代出版社 [Xiandai Chubanshe].
- Zhao, Pingan [赵平安] (2009). 隶变研究 [Research on the Li-Variation]. Baoding: 河北大学出版社 [Hebei Daxue Chubanshe].

On the Origin of Arabic Script

Kamal Mansour

Abstract. For the past two centuries, scholars have debated the origin of Arabic script, the youngest of the Semitic scripts. While one camp pointed to Nabatean as the sure ancestor, another favored Syriac instead. By examining the each ancestor visually and historically, one finds evidence for each point of view. Is it reasonable to insist on a single ancestor for Arabic script? The historical examples of Proto-Sinaitic and Ugaritic scripts demonstrate that a single script can be shown to have features amalgamated from more than one source. Detailed examination of the features of early Arabic script leads us to conclude that both Nabatean and Syriac strongly influenced its development. Finally, we demonstrate that particular details of cursive linking in Arabic script replicate analogous behavior in Syriac.

The origin of Arabic script has been much discussed and disputed in the last two centuries. Scholarly opinion is divided primarily into two camps: one says Arabic script descends from Nabatean, while the other points at Syriac. In the 9th century, the Arab historian, al-Baladhuri, recounted that three men from the tribe of Ṭayy had fashioned Arabic script “in a manner like Syriac” (Al-Baladhuri, 1969). About one thousand years later in 1865, orientalist T. Nöldeke published his study which concluded that Arabic writing descended from Nabatean script (Grohmann, 1971). About one hundred years later, semiticist J. Starcky argued in favor of Syriac because of its structural resemblance to Arabic script (Starcky, 1966). In 1993, arabist B. Gründler published her doctoral work at Harvard University in which she collected exhaustive material to demonstrate a gradual progression from Nabatean writing to early Arabic writing (Gründler, 1993). This publication displays the variety of glyph forms for each letter of the Nabatean alphabet in its long transition to Arabic script (Fig. 1). It is interesting to note that when Gründler later wrote the section on Arabic script in the *Encyclopedia of the Qurʾan* (Gründler, 2001), she stated that Arabic writing was also likely

Kamal Mansour
Monotype
kamal.mansour@me.com

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1.
Fluxus Editions, Brest, 2019, p. 245–255. <https://doi.org/10.36824/2018-graf-mans>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

influenced by Syriac calligraphy. A few years later, Gründler stated that "...the Nabatean script (attested 100 BCE-350 CE) was the genetic ancestor of the current Arabic alphabet" (Gründler, 2006).

In 1997, semiticist F. Briquel-Chatonnet entered the discourse in favor of Syriac, arguing strongly against the primary use of individual glyph shapes to demonstrate a relationship between scripts (Briquel-Chatonnet, 1997). She asserted that writing systems need to be compared by their overall look on the page, while also taking into consideration historical and cultural factors such as the status and prestige associated with each script. A script must essentially be examined as a whole system and not merely as a collection of glyphs. Briquel-Chatonnet points out some important visual differences between Nabatean and Syriac. In terms of alignment, Nabatean characters appear to be suspended from a common horizontal line, with the lower part of the letters uneven. On the other hand, Syriac letters are sitting on a common base line, along which the letters are also connected. In terms of proportions, Nabatean characters can be characterized as being taller than wider, while the Syriac letters are mostly wide with a few tall strokes here and there. Figure 2 demonstrates the visual contrast between the two scripts by showing a Nabatean papyrus from the 2nd century CE (Starcky, 1954) opposite a Syriac parchment in informal style from the 3rd century CE (Teixidor, 1990).

When it comes to scholarly discourse about the origins of a script, the use of terms such as "genetic," "descendant," and "ancestor" imply a sole family-line view of each script. But, can't the traits of a script be adopted from more than one source? Ugaritic script is a superlative example of a hybrid script developed by deriving traits from various scripts and amalgamating them (Fig. 3). Its inventor adopted its phonetic repertoire of 27 consonants, as well as their alphabetic order, from a similar Semitic language, while developing their shapes using components of Mesopotamian cuneiform writing (Pardee, 2012). Ugaritic script cannot be called the descendant of a sole script, but we can see that its elements evidently hark back to at least two sources. Although many scripts have slowly undergone changes over a long period of time, a few—such as Ugaritic—were created in a relatively short time through the deliberate mixing of traits. Proto-Sinaitic script, the first consonantal alphabet, falls also into that category; its inventor borrowed from the shapes of Egyptian characters, while naming the resulting letters to reflect their phonetic values in a Semitic language—a brilliant amalgamation of traits. Also, we should not neglect the fact that this inventor had to first identify all the consonants of the language, which in itself is a grand feat of linguistic insight.

With regard to the basic letter shapes of early Arabic writing, J.F. Healey has amply demonstrated that many of them show similarity to both Nabatean and Syriac shapes (Healey, 2000), although sev-

eral cannot be readily derived from Syriac. Such an observation should not be surprising because both scripts ultimately derive from Imperial Aramaic, albeit through different paths.

L. Nehmé has written extensively about more recently discovered samples of a script she dubs “Nabateo-Arabic” that is “neither Nabatean nor Arabic but somewhere on its way between the two” (Nehmé, 2017). Dating to the 4th and 5th centuries, these writings show “some letters [that] are more still reminiscent of their equivalent in calligraphic Nabatean or are still ‘on their way’ to Arabic”. Nehmé further asserts that “sometime between the end the 5th century and the end of the 6th century,” the forms of Arabic letters were standardized, possibly through the influence of Syriac writing.

In their early form, many of the Arabic letters were polyvalent; i.e., each letter could represent several sounds. For instance, the letter Jīm—bearing no dots as in modern orthography—represented [ǧ], [ḥ], or [ḫ]. Such polyvalence clearly made reading more difficult since the reader had to determine the appropriate sound for each letter based on the word, together with the broader context. More importantly, it demonstrates that Arabic writing—at the time—was based on another alphabet that supported fewer consonants. The 22-letter repertoire of Nabatean was based on that of Aramaic, while Arabic possessed 28 consonants. In order to represent 28 phonemes with 22 letters, some ambiguity was bound to result. With a repertoire of 23 letters, Syriac would not have done much better. On the other hand, had the Arabic repertoire been based on the ample set of 29 Ancient South Arabian consonants, it would have been an adequate fit (Nebes and Stein, 2004). However, the paths of these two scripts do not seem to have crossed.

One emblematic trait of Syriac writing does not appear to have been mentioned by the literature related to the origin of Arabic script. Syriac writing is cursive in the sense that all letters of a word link to each other along the horizontal base line. While this behavior is true in general, a subset of the letters does not conform to it. Of the 23 letters of the Syriac alphabet, eight link to their neighbor on the right, but never link on the left, even in mid word. This set of eight consists of the following letters: Alaf (ʿ), Dalat (d), He (h), Waw (w), Sade (š), Zayn (z), Rish (r), Taw (t).

In Arabic, five of the eight phonetically equivalent letters demonstrate this same linking behavior: Alef (ʿ), Dal (d), Reh (r), Zayn (z), Waw (w). One might puzzle, was this behavior borrowed from Syriac into Arabic, or the contrary? In either case, it seems most unlikely that such an unusual pattern common to two scripts would have come about accidentally. The written record shows partial evidence of such behavior in Syriac as far back as the 3rd century CE (Teixidor, 1990). By the time the Syriac codex manuscripts appeared in the 5th century, the formal *estrangela* style of Syriac had matured and become standardized. G. Kiraz has demonstrated that by then, this linking behavior had become

the norm, placing it long before Arabic script had reached full development (Kiraz, 2012). Figure 4 shows examples of this linking behavior in a Syriac manuscript (dated 464 CE), while figures 5 and 6 shows parallel examples in early Arabic inscriptions, one dating to 568 CE, and the other to 677 CE.

Is it not possible that both Nabatean and Syriac contributed to the formation of Arabic script at various stages? Is the discourse about script origin perhaps too steeped in assumptions that hinder an objective examination of the subject? In *The Shape of Script*, R. Salomon describes the slow, gradual changes that a script can undergo as a “constant natural process of evolution” (Salomon, 2012). Among other terms commonly used in the context of script are “descendant,” “ancestor,” and “genetic”—all biological terms. The terms we use certainly have an influence on our thinking. In real life, we know that a cat cannot be crossbred with a horse. Might we be unwittingly extending this type of reasoning to scripts? And yet, we know that scripts are *not* living entities. They are symbolic systems invented by human minds primarily to keep records and to represent spoken language. The transformation of a script, as observed over time, can resemble the slow adaptations of living beings, but in reality, they vary visually only as a result of the human tendency to introduce change. There is nothing “natural”—in the biological sense—about the changes that a script undergoes. At each stage, humans slowly vary the shapes that they write, gradually resulting in long-term changes.

In conclusion, we must weigh the evidence regarding the origin of Arabic script.

With the body of evidence on each side—i.e., Nabatean and Syriac, the scales do not tip readily in favor of a single, exclusive source for Arabic script. The alphabetic repertoire of Arabic script is evidently of Nabatean origin, while at some later time, its letter shapes and its connecting behavior were probably influenced by Syriac. The inscription at Zabad (Fig. 7) is written in three languages (Syriac, Greek, and Arabic), indicating the close coexistence of multiple scripts in the Levant (Grohmann, 1971). It is most reasonable then to conclude that the traits of Arabic script have at least two sources, Nabatean and Syriac.

References

- Al-Baladhuri, Ahmad (1969). *The Origins of the Islamic State, Being a Translation from the Arabic Accompanied with Annotations Geographic and Historic of the Kitâb Futuḥ Al-Buldân by F.C. Murgotten. Part II*. New York: AMS Press.

- Briquel-Chatonnet, Françoise (1997). “De l’araméen à l’arabe: quelques réflexions sur la genèse de l’écriture arabe”. In: *Scribes et manuscrits du Moyen-Orient*. Ed. by François Déroche and Francis Richard. Paris: Bibliothèque nationale de France.
- Grohmann, Adolph (1971). *Arabische Paläographie*. Vol. 94. Vienna: Österreichische Akademie der Wissenschaften.
- Gründler, Beatrice (1993). *The Development of the Arabic Scripts: From the Nabatean Era to the First Islamic Century according to Dated Texts*. Vol. 43. Atlanta: Scholars Press.
- (2001). “Arabic Script”. In: *Encyclopedia of the Qur’an*. Ed. by J. D. McAuliffe. Vol. 1. Leiden: Brill, pp. 135–44.
- (2006). “Arabic Alphabet: Origin”. In: *Encyclopedia of Arabic Language and Linguistics*. Ed. by Lutz Edzard and Rudolf de Jong. Leiden, Boston: Brill.
- Healey, John F. (2000). “The Early History of the Syriac Script—a Reassessment”. In: *Journal of Semitic Studies* 45, pp. 55–68.
- Kiraz, George (2012). “Old Syriac Graphotactics”. In: *Journal of Semitic Studies* 57, pp. 231–264.
- Nebes, Norbert and Peter Stein (2004). “Ancient South Arabian”. In: *The Cambridge Encyclopedia of the World’s Ancient Languages*. Ed. by R.D. Woodward. Cambridge: Cambridge University Press.
- Nehmé, Laïla (2017). “New Dated Inscriptions (Nabataean and Pre-Islamic Arabic) from a Site Near Al-Jawf, Ancient Dūmah, Saudi Arabia”. In: *Arabian Epigraphic Notes* 3, pp. 121–164.
- Pardee, Dennis (2012). *The Ugaritic Texts and the Origins of West Semitic Literary Composition*. Oxford: Oxford University Press.
- Salomon, Richard (2012). “Some Principles and Patterns of Script Change”. In: *The Shape of Script: How and Why Writing Systems Change*. Santa Fe: School for Advanced Research Press.
- Starcky, Jean (1954). “Un contrat nabatéen sur papyrus”. In: *Revue Biblique*, pp. 161–181.
- (1966). “Pétra et la Nabatène”. In: *Dictionnaire de la Bible: Supplément VII*. Paris: Letouzey et Ané, pp. 886–1017.
- Teixidor, Javier (1990). “Deux documents syriaques du III^e siècle ap. J.-C., provenant du Moyen Euphrate”. In: *Comptes-rendus des séances de l’Académie des Inscriptions et Belles-Lettres* 134, pp. 144–166.

Handwritten text in a cursive script, likely a historical document or manuscript. The text is arranged in approximately 15 horizontal lines. The script is dense and difficult to decipher due to its cursive nature and the image's quality. The text appears to be written in a historical form of a European script, possibly Latin or Italian. The final line of the document includes the number "73" written in a larger, bolder hand, indicating the page number.

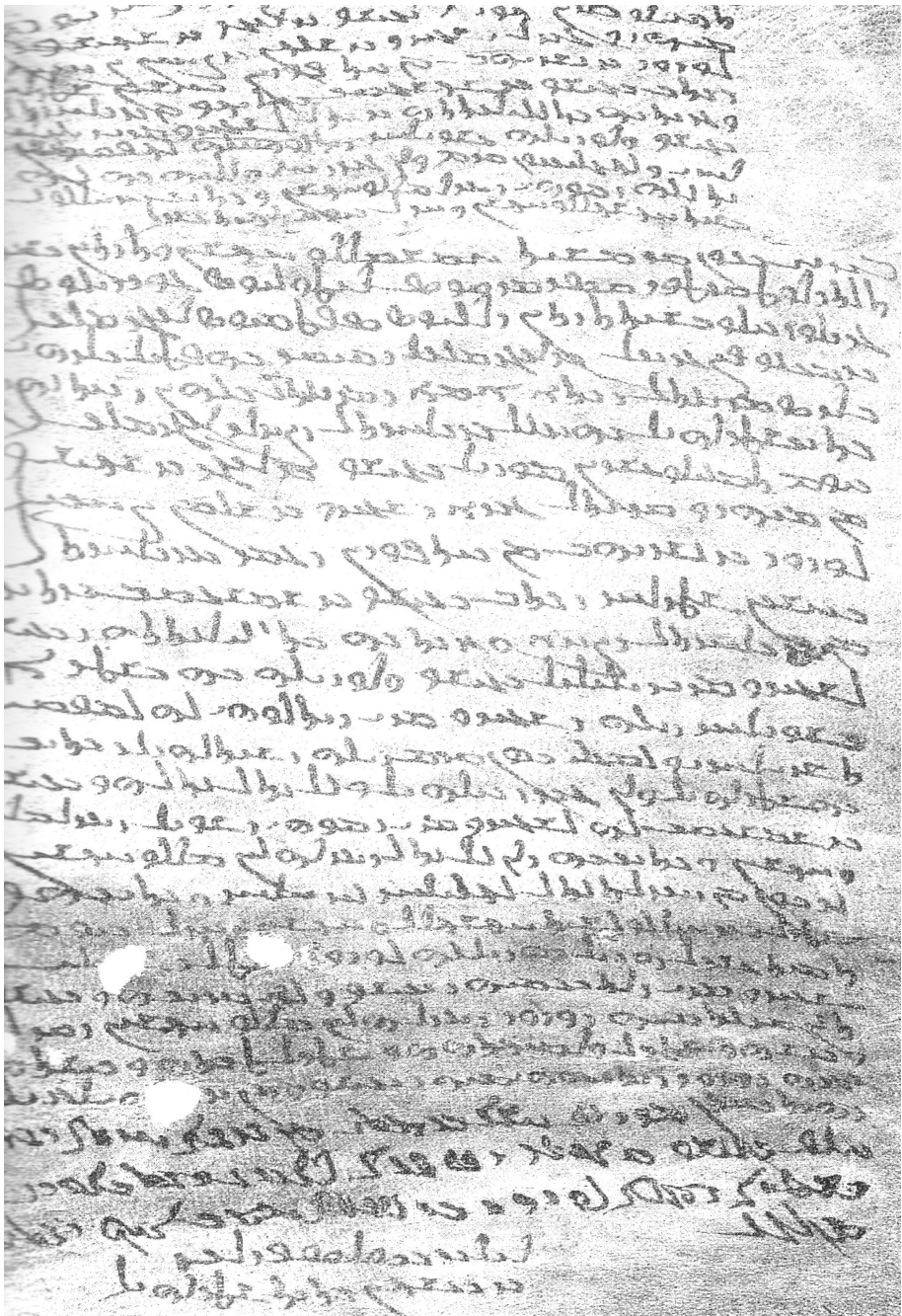


FIGURE 1. Visual contrast of two early manuscripts: Nabatean (previous page), Syriac (this page)

	Left Connection	Right & Left Connection	Right Connection	No Connection
N1				𐤌
N2	𐤌𐤍	𐤌	𐤌	
N3	𐤌		𐤌	𐤌
N4	𐤌𐤍		𐤌	𐤌
N5				𐤌𐤍
N6	𐤌𐤍	𐤌	𐤌𐤍	𐤌
N7	𐤌𐤍	𐤌	𐤌𐤍	𐤌
N8	𐤌𐤍	𐤌	𐤌	𐤌
N9	𐤌𐤍	𐤌		
N10	𐤌𐤍		𐤌𐤍	𐤌
N11			𐤌𐤍	𐤌
N12			𐤌	𐤌
N13	𐤌		𐤌𐤍	𐤌
N14	𐤌𐤍			
N15			𐤌𐤍	𐤌
N16	𐤌𐤍			𐤌
N17	𐤌𐤍		𐤌	
N18	𐤌	𐤌		
N19	𐤌𐤍	𐤌	𐤌	𐤌
N20	𐤌𐤍	𐤌	𐤌	𐤌
N21	𐤌𐤍	𐤌	𐤌	𐤌

م

	Left Connection	Right & Left Connection	Right Connection	No Connection
A1			م	م م م
A2	م م			
A3	م م	م م	م	
A4	م	م		
A5	م	م م	م	

FIGURE 2. Samples of the letter Mīm, from Nabatean to early Arabic (Gründler, 1993)

ʔa	b	g	ḥ (x)	d	h
w	z	ḥ (ḥ)	ṯ	y	k
š	l	m	ḏ (ð)	n	z (θ)
s	ṣ	p	ṣ	q	r
ṯ (θ)	ḡ (y)	t	ʔi	ʔu	s ₂



FIGURE 3. An abecedarium of Ugaritic script

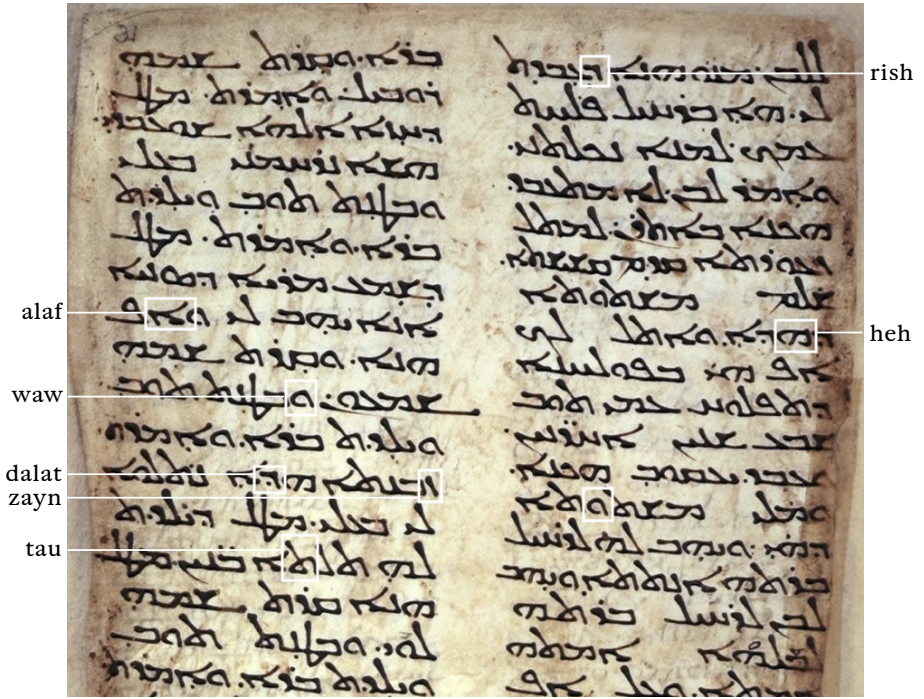


FIGURE 4. Examples of Syriac letters that do not link to their left neighbor, British Library (Add MS 14425)

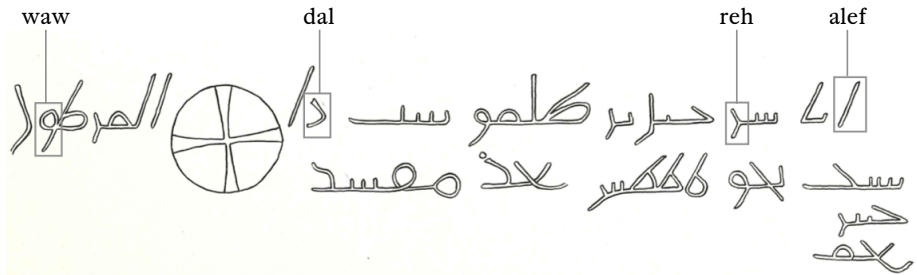


FIGURE 5. Examples of Arabic letters that do not link to their left neighbor, Inscription from near Harrân (Southeastern Turkey) dated 568 CE

On the Writing System of Arabic: The Semiographic Principle as Reflected in Nashī Letter Shapes

Joseph Dichy

Abstract. Arabic writing as we know it was codified between the 1st/7th and 3rd/9th centuries. It is much more ancient and had appeared with another type of letter shapes akin to South Arabian writings around 800 years BC (Robin, 1991). The script borrowed the basis of its letter shapes from Nabataean writing (Ba‘albaki, 1981; Healey and Smith, 2009). Arabic script became more rational and regular due to what S. Aurox called, with reference to the French language, *grammatization processes*, which can be illustrated with Arabic, due to the very rich medieval sciences literature that was contemporary or immediately subsequent to its codification.

This paper is concerned with one aspect of the grammatisation process, that of letter shapes in the *Nashī* style of calligraphy, which was codified by Ibn Muqla (4th/10th cent.) and his followers. The paper presents the basic drawings included in the building of letters as designed by Ibn Muqla. It also highlights the fact that the relatively small number of shapes in Arabic is due to the cursive line inherited from Nabatean script. The graphic word-form with its complex set of morphological structures (extensively described in Dichy 1990a; 1997a), was another inheritance of earlier Semitic writing tradition (which isolated word-forms since Phoenician writing). Word-form structure resulted in a special style of final letters shapes, which exist in Hebrew with five letters, but were systematized in Arabic in a way that produced a basic opposition between initial and medial letter-shapes on the one hand, and final (end-of-word) shapes on the other. The result, including a few letters that escaped the opposition, is also presented.

The overall view highlights, in the author’s theory of writing (Dichy, 2019), the way in which the *Semiographic Principle* parallels the *Phonographic Principle* in the writing of the Arabic language. This approach considers writing systems as analytic (Dichy, 2017). This means that the writing system can be considered as a collective cognitive analysis of the oral language. Analytic results are then projected on a writing support according to the semiographic structures of the script. The orthography of the language features the complex relations between the two principles. One point underlying the present paper is that the collective cognitive developments associated with a writing system are related to gramma-

Joseph Dichy
Professor of Arabic Linguistics Lyon - France
joseph.dichy@yahoo.fr

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1.
Fluxus Editions, Brest, 2019, p. 257–268. <https://doi.org/10.36824/2018-graf-dich>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

tisation processes, which occur in a given historic period (there can therefore be successive grammatisation periods in a given language and culture).

1. Introduction: The Analytic Approach to the Emergence of a Writing System

This contribution follows previous works on the writing system of Arabic (based on Dichy 1990a,b) and the development of a cognitive approach to the emergence of a given script, which gives birth to a writing system through a metalinguistic process, the first insights of which can be found in Condillac's works (1746; 1775), and in the 20th century, in those of Vygotsky (1934). This overall view has been recently presented in Dichy (2017; 2019) in a more elaborate version, and can only be hinted at here. In very short words, the metalinguistic process of scriptural analytics operates along two interwoven lines:

- *Analytic operations*—the object of which is the oral language at phonetic, morphological and morphotactic levels—, result in *segmentation* and *identification* processes based on what I have described as the *Inventory Principle (IP)*. (In alphabetic writings, the Inventory Principle is cenic, with a number of pleremic adjustments.)
- *Synthetic operations* project the result of analytic operations on writing supports (stone, parchment, paper, and nowadays screen) according to the conventions proper to the *semiographic characterisation* of a given language/script, i.e., according to the *Semiographic Principle (SP)*.

This paper is concerned with one of the aspects of the metalinguistic processes that lead to the emergence of a writing system, namely, the shapes of letters considered as the result of synthetic operations leading to the semiographic characterisation of Arabic. This is related to the visual characterisation of both letters and words.

1.1. Short Recall of a Few Historical Facts about the Arabic Writing System

We leave it, with great relief and pleasure, to our Colleague Dr. Kamal Mansour to present, in this conference (Mansour, in this volume), the origins of Arabic Script, which is the latest of the Semitic scripts.

Written Arabic is in fact much more ancient than the writing we currently know and had appeared with another type of letter shapes

around the early 1st millennium BC¹. The older script was borrowed from South Arabian writings. The writing in use to-day, the earliest form of which seems to go back to the 4th century CE for the least—i.e., to some 200 years before Islam—borrowed the basis of its letter shapes from Nabatean writing (Ba‘albakī 1981; Healey and Smith 2009). This entailed a good deal of adaptation, because the Syriac writing from which Arabic script was borrowed only had 22 graphemes (Robin 1991; Ba‘albakī 1981; Dichy 1990b).

The emergence of the latter into a codified writing system is related to considerable historiographic and linguistic developments that give a significant amount of information on its early traditions and progresses (Dichy 1990a,b; Abbott 1939b). On the other hand, the art of writing Arabic—which parallels that of good saying—has known remarkable developments since the beginning of Islam. The 4th/10th century² vizier Ibn Muqla (d. in 940) has authored, on the basis of his own experience as a calligrapher, a famous epistle on writing, which has been a basis for later Arabic calligraphy³, and will be used extensively in the main sections of this contribution.

The emergence of Islamic culture was related to heavy needs in written documents. In addition to the Qur’an and the Ḥadīṭ (the teachings of the Prophet of Islam), one witnessed great development in medieval sciences (mathematics, astronomy, medicine and surgery, logics, grammar and lexicography, religious sciences..., etc.), as well as many translations (mainly from Syriac and Greek). On the other hand, the emerging Islamic culture needed to ensure knowledge and transmission of the language of what we would call to-day the linguistic community of the Prophet of Islam. Basra and Kufa scholars, living in Iraq, endeavoured to compile the corpus of the poetry of the ancient Arabs in *diwān-s* (poetic records). The compilation ensured contextualisation of the words found in the Koran and Ḥadīṭ, and a far-reaching knowledge of the language. In addition, words were extensively recorded in Arabic dictionaries from al-Ḥalīl (d. ar. 175 H/792 G) onwards. It is to be noted that, in addition to his role in the devising of the diacritic representation of vowels (Abbott, 1939a), al-Ḥalīl elaborated a method for the generation of

1. Ch. Robin has highlighted the fact, based on recently discovered inscriptions going back as far as the 8th century BC, that Arabic had been written using a South-Arabian alphabet, which was more adapted to its structure than the Syriac script, which was later borrowed (Robin, 1991, pp. 127–129).

2. Traditionally, the first date is that of the Islamic calendar (‘H’ for Hegira) and the second is that of common Gregorian years (symbol ‘G’).

3. On Ibn Muqla, see Abbott (1939a), and Osborn (2017, pp. 15–16). The first eight pages of Ibn Muqla’s Epistle are reproduced *in fac simile* (after the 1663 manuscript of the Cairo National Library) with a commentary, in Massoudy (1981, pp. 40–41). One finds many quotations of Ibn Muqla in al-Qalqaṣandī (vol. 3, p. 23 onwards, chapters on writing, letter shapes and calligraphy).

virtual ‘constructs’ (in our terms, ‘roots’) that allowed building exhaustive dictionaries of the language of the Ancient Arabs, and gave birth to the first dictionary aiming at including the whole lexicon of a given language in the history of mankind (Dichy, 2014). An elaborate and reliable script was obviously the basis of the emerging written transmission of knowledge, which paralleled, during the first centuries oral transmission (Schoeler, 2002).

An essential development complemented the emergence of the Arabic script: conventions were needed in order to stabilise the reading of the Koranic text. In the 2nd/8th century, al-Ḥalil devised a system of very comprehensive secondary diacritics (Abbott 1939b; Dichy 1990b), which is still in use to-day.

1.2. Illustrating the Processes That Lead to One Crucial Aspect of the Writing System

How did the writing system of Arabic emerge? To this question, one can only give conjectural answers. The fact is, writing became more rational and regular after its first period, due to what Auroux (1998; 2010) called, with reference to the French language, grammatization processes. This hypothesis can be illustrated with Arabic, on the basis of descriptive texts, due to the later appearance of its writing.

In this paper we will consider, from a descriptive standpoint, the shapes of Arabic letters and their development into a graphic system that differentiated analytically between graphemes representing phonemes, in accordance to what I described as the *Phonographic Principle*. This was not achieved immediately, as will be illustrated. Visual shapes of letters are differentiated according to the *Semiographic Principle*. In the 4th/10th century, the *Nashī* style of letters was systematically organized by the vizier and great calligrapher Ibn Muqla.

The graphic word-form is a complementary aspect to the shape of letters (Dichy, 1990b; 1997a). Its complex set of morphological structures occurred at a very early stage, because of long time Semitic writing traditions, which isolated word-forms in various ways since the Phoenician writing (dots separating words and special end-of-word shapes of a few letters, which appeared in early Hebraic writing...). Nevertheless, this second aspect cannot be fully developed here for lack of time and space. The figure below hints at the question of the structure of the Arabic word-form. Another aspect will be presented in some detail in relation with the shapes of letters: the word-form structure resulted in a special style of final letters shapes, which was systematized in Arabic.

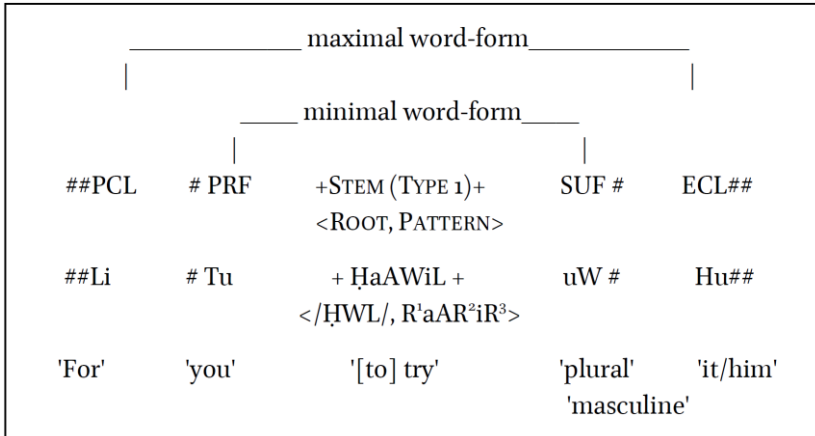


FIGURE 1. Recall of the structure of the word-form (from Dichy 1997a)

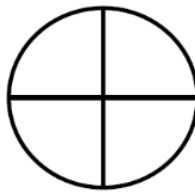


FIGURE 2. Ibn Muqla's virtual circle

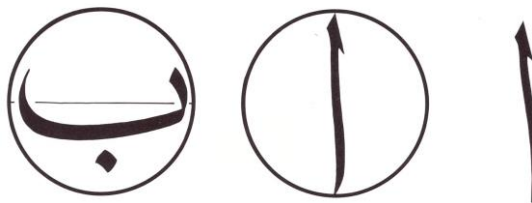


FIGURE 3. An intuitively drawn alif determines the height of the vertical diameter, which in turn determines the horizontal dimensions of a final bā' (from Massoudy 1981, p. 38)

2. Ibn Muqla's Epistle and the Coding of Writing on the Basis of the Nashī Style

Ibn Muqla is considered by medieval biographical sources as the first calligrapher who put down a codified set of regulations for letter shapes in Arabic. In this code, the dimensions of letters are reckoned after the writing of a letter *ʿalif* included in a circle featuring two perpendicular diameters.

The length of the *ʿalif* is that of the vertical diameter. The letter *bā'* covers the horizontal one accordingly, as shown in the figure below.

Later calligraphers have added a convention according to which the dimensions of letters are measured by dots. Ibn Muqla also broke down the movements of the pen (*qalam*, borrowed from the Greek *kalamos*) into five fundamental strokes, which H. Massoudy⁴ takes up in the following explanatory drawings:

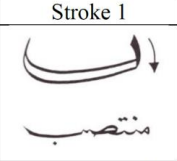
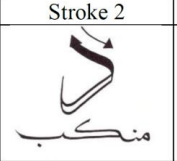
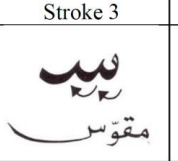
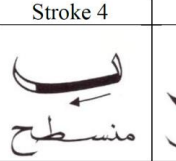
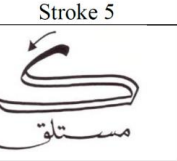
Stroke 1	Stroke 2	Stroke 3	Stroke 4	Stroke 5
				
Erect stroke (ES)	Inclined stroke (IS)	Curved stroke (CS)	Flattened stroke (FS)	Thrown down stroke (TS)

FIGURE 4. Ibn Muqla's five basic strokes (reproduced from Massoudy 1981, p. 39)

These basic movements of the pen are then developed through a description of the drawing of every letter. These can be found with more details than in Ibn Muqla's epistle in the 3rd volume of al-Qalqaṣandī's encyclopaedic work *Ṣubḥ al-aṣṣā fī ṣinā'ati l-inṣā*, in the chapters on writing, letter shapes and calligraphy (vol. 3: 23 onwards). In the coding of the shapes of letters, the Nashī style of writing thus became a reference style, which remains globally true to-day. The very first printed edition of the Qurʾān, due to the Būlāq Printing press in 1923, was realised in Nashī characters.

The reference status of this style, and the coding of Arabic calligraphy by Ibn Muqla and the following traditions allows considering Nashī letter shapes as the basis of what I have called *semiographic characterisation*, which I will now endeavour to describe with regards to letters.

4. Hassan Massoudy is a well-known iraqi calligrapher living in Paris. See <http://hassan.massoudy.pagesperso-orange.fr/>.

The choice of the Nashī writing style as a basis for the study of the semiographic characterisation of the Arabic script could of course be discussed: other candidates could be proposed, such as the Kūfī style, which remained in use for Qurʾan manuscripts for a few centuries even after Ibn Muqla's referential coding (Deroche, 2004); or the Maḡribī style, which has been analysed from a visual standpoint by Lakhdar Ghazal, who proposed reading betterments. These of course should be the subject of further studies, which should, in my opinion, benefit by the analyses presented here.

3. The Constrained System of Letter Shapes in Arabic and Word-Form Visual Identification

The letter shapes of Arabic are constrained by a cursive line, which means that *the number of drawings is limited*, and cannot exceed a few figures. The same is true of Syriac writings from which Arabic script, as we know, has been borrowed. Syriac had 22 letters, whereas Arabic includes 29 letters or 28, depending on alphabetic conventions⁵. The 'ancestor' of Arabic script therefore required a number of adaptations. This situation provides an explanation for two general features of Arabic writing (Dichy, 1990b):

- I have described diacritical dots, the width of which is the same as that of the cursive line, as *primary diacritics*. These actually belong to the letter—as opposed to *secondary diacritics* used for short vowels, undetermined case endings and the doubling of a consonant, in addition to diacritic *hamzas*. The width of secondary diacritics is much thinner, because they are drawn with the tip of the pen.
- In order to complement primary diacritical dots, the script resorts to *breaking the cursive line* to identify six letters. These breaks used to be called mnemotechnically, in bygone days, by people working in printing presses, “the *ḍwār* letters”: دوار. One needs to add to these four letters ḍ and j, which feature an additional diacritical dot.

5. The alphabet, as it is nowadays taught in schools, includes 28 letters. From medieval times until the 1950s, it was considered as comprehending 29 letters, the first letter *ʿalif* being a *hamza*, in accordance with the acrophony principle. In order to represent the long vowel *ā*, a *lam-alif* was traditionally added at the end of the alphabet between *wāw* and *yāʾ*. The double letter *lam-alif* was used because the long vowel *ā* could not be started with. (These point were outlined in the 4th/10th century by Ibn Jinnī, *Sirr Ṣināʿat al-ʿirāb*, vol. I: 41). The omission of the *lam-alif* in the current teaching of the alphabet in schools does not only reduce the number of letters from 29 to 28, it also affects the previous series of *muʿtall* letters (i.e., “weak” or subject to transformation phonemes) positioned at the end of the alphabet. We sincerely hope one of the Ministers of Education will read Ibn Jinnī and restore the age-old traditional Arabic alphabet.

The counterpart of letters in most writings is the word-form. I have recalled above the fact that Semitic writings—among other scriptural traditions—indicated *word boundaries* by a dot, a stroke or the end of a line. Hebraic writing did in addition, as is still the case in Modern Hebrew, have special end-of-the-word shapes for five letters. It is also the case of eight letters in Syriac. This phenomenon has become a mainstream feature in Arabic, where the style of letter has come to oppose two basic letter shapes:

initial or mid-word shape *vs.* end-of-the-word shape.

Nevertheless, end-of-word letter shapes cannot be considered as a strictly enforced principle, because it encounters a few practical impediments, most of which are related to the breaking of the cursive line by some letters. Let us now consider the basic forms of letters.

4. Basic and Initial Letter Shapes

End-of-word forms are analysed in §4.3. Under the constraint of cursive writing, initial letter shapes are limited to two basic models: *the stroke* and *the rounded form*. The first group of letters can be described on the whole with the five strokes identified by the vizier Ibn Muqla. The second one is introduced in this presentation:

- The first group develops the Erect stroke (ES), in combination, when needed, with
 - the Leaning stroke (LS),
 - the Curved stroke (CS),
 - the Thrown down stroke (TS),
 - a combination of two or more of these strokes,
 - the breaking of the cursive line.
- The second group introduces round shapes (RS), in combination, in one case with a breaking of the cursive line.

These two groups allow the building of a taxonomy, which covers all the letters of the alphabet, every item being strictly included in a single class. This can be seen in the tables below. Final shapes of letters are also divided into two parts (§4.3).

In the figures 5 to 7 below, letters are presented between two small lines, showing whether it goes below the cursive line, or above.

5. End-of-Word Shapes

Letters are presented in Fig. 7 both attached to the cursive line and in ‘isolated’ form, the latter being liable to be drawn or printed slightly

	Single stroke letters			Two-strokes letters		Three-strokes letters		
	Simple Erect stroke	Triple Curved stroke	Vertically prolonged stroke	Inclined + Flattened strokes	Inverted inclined + Flattened strokes	Inclined + Thrown down + Flattened strokes	Inclined (horizontal) + Flattened + Simple Erect strokes	Inclined (horizontal) + Flattened + Vertical prolonged strokes
Letters in initial and middle-of-word form	ب - ت - ث - ذ - ي -	س - ش -	ل -	ج - ح - خ -	ع - غ -	ك -	ص - ض -	ط - ظ -
Letters with cursive line breaking	ر - ز -		ا -	د - ذ -				
Letters with a different mid-word shape					ف -			

FIGURE 5. The basic model is that of the different types of strokes. It covers 23 letters

	Single rounded form letters		Curved inclined stroke + rounded form letters
	Simple rounded form	Simple rounded form, inverted drawing	A curved inclined stroke is continued into a rounded form
Letters in initial and middle-of-word form	ف - ق -		ه -
Letters with cursive line breaking	و -	م -	
Letters with a different mid-word shape			ق -

FIGURE 6. The rounded form model only concerns 5 letters

higher (ex.: ن - ن) and/or in a somewhat different way than the former; examples of small differences : ك - ك / ي - ي; bigger differences can be found in the forms of initial, mid-word and end-of-word shapes of the letters ع and ه.

The above shapes visually present readers with the left-hand word boundary, i.e., with the end of the word-form.

This is not the case with the دوار group of letters (ا, و, ز, د), which break the cursive line, for obvious reasons (the final and the mid-word or initial form of the letter remains the same). The existence of these letters is partly responsible for the fact that letter-shapes indicate word boundaries most of the time, but not always.

		Final shapes in contrast (in Nashī writing)
The ب model	ب - ب / ت - ت / ث - ث	ف - ف /
The ن model	ن - ن / ل - ل س - س / ش - ش / ص - ص / ض - ض	ق - ق /
The ي model	ي - ي / ح - ح	
The ك model	ك - ك	
The ع model	ع - ع / غ - غ / ج - ج / ح - ح / خ - خ	
The ط model	ظ - ظ	
The م model	م - م	
The ه model	ه - ه / ة - ة	

FIGURE 7. Final shape of letters in letters featuring no cursive line breaking

6. Conclusion

The *Semiographic Principle* illustrated above is related to a theoretical approach which considers writing systems as *analytic* (Dichy 2017; 2019). This means that the writing system can be considered as a collective cognitive analysis of the oral language. We have endeavoured here to illustrate the way in which the result of this analysis is projected on paper semiographically, i.e., in a systematic way in which letter shapes are clearly contrasted. The above elaboration owes a lot to the vizier Ibn Muqla's systematic encoding. It emphasises in addition the role of final letter shapes in the identification of word-forms and in the reading process.

References

- Abbott, Nabia (1939a). "The Contribution of Ibn Muklah to the North Arabic Script". In: *American Journal of Semitic Languages and Literatures* 56, pp. 70–83.
- (1939b). *The Rise of the North Arabic Script and Its Kur'anic Development, with a Full Description of the Qur'ān Manuscripts in the Oriental Institute*. Chicago: The University of Chicago Press.
- Atanasiu, Vlad (1999). *De la fréquence des lettres et de son influence en calligraphie arabe*. Paris, Montréal: l'Harmattan.
- Auroux, Sylvain (1998). *Le langage, la raison et les normes*. Paris: Presses Universitaires de France.

- (2010). “Sur le mythe de la langue”. In: *Pour la (socio)linguistique. Pour Louis-Jean Calvet*. Ed. by Médéric Gasquet-Cyrus et al. Paris: L’Harmattan.
- Ba‘albakī, Ramzī [بعليكي، رمزي] (1981). الكتابة العربية والسامية. [Arabic and Semitic Writing]. Beirut: دار العلم للملايين [Dār al-‘ilm li-l-malāyīn].
- Condillac, E. Bonnot de (1746). “Essai sur l’origine des connaissances humaines”. In: *Œuvres philosophiques de Condillac*. Ed. by G. Le Roy. Paris: Presses Universitaires de France.
- (1775). “Grammaire”. In: *Œuvres philosophiques de Condillac*. Ed. by G. Le Roy. Paris: Presses Universitaires de France.
- Deroche, François (2004). *Le livre manuscrit arabe. Préludes à une histoire*. Paris: Bnf.
- Dichy, Joseph (1990a). “Grammatologie de l’arabe I. Les sens du mot harf, ou le labyrinthe d’une évidence”. In: *Studies in the History of Arabic Grammar II*. Ed. by K. Versteegh and M.G. Carter. Amsterdam, Philadelphia: Benjamins, pp. 111–128.
- (1990b). “L’écriture dans la représentation de la langue: la lettre et le mot en arabe”. Doctorat d’État. Université Lumière Lyon 2.
- (1997a). “Deux grands “mythes scientifiques” relatifs au système d’écriture de l’arabe”. In: *L’Arabisant* 32–33, pp. 67–86.
- (1997b). “Pour une lexicomatique de l’arabe: l’unité lexicale simple et l’inventaire fini des spécificateurs du domaine du mot”. In: *Meta* 42, pp. 291–306.
- (2014). “Al-Ḥalīl’s Conjecture: How the First Comprehensive Dictionary in History Was Invented”. In: *Arab and Arabic Linguistics: Traditional and New Theoretical Approaches*. Ed. by Manuella Giolfo. Oxford: Oxford University Press, pp. 39–64.
- (2017). “The Analytics of Writing, Exemplified by Arabic, the Youngest of the Semitic Scripts”. In: *Approaches to the History and Dialectology of Arabic, in Honour of Pierre Larcher*. Ed. by M. Sartori, M.E. Giolfo, and Ph. Cassuto. Leiden, Boston: Brill, pp. 29–56.
- (2019). *Études sur le système d’écriture et la linguistique de l’écrit en arabe*. Villeurbanne (Lyon): Aradic-Monde arabe.
- Driver, Godfrey R. (1976). *Semitic Writing: From Pictograph to Alphabet*. 3rd ed. Oxford: Oxford University Press.
- Grainger, Jonathan, Joseph Dichy, et al. (2003). “Approche expérimentale de la reconnaissance du mot écrit en arabe”. In: *Faits de langue* 22, pp. 77–86.
- Healey, John F. and G. Rex Smith (2009). *A Brief Introduction to the Arabic Alphabet: Its Origins and Various Forms*. London, San Francisco, Beirut: Saqi.
- Ibn al-Ḥāğib [ابن الحاجب] (1975). الشافية في علي التصريف والخط. [Commentary on the Satisfying [Book] on the Two Sciences of Morphology and Writ-

- ing]”. In: شرح الشافية، الأستراياضي، [Commentary on the *Shāfiya*]. Ed. by Nūr al-Ḥasan [الحسن, نور]. Beirut: دار الكتب العلمية [Dār al-kutub al-‘ilmiyya].
- Jaffré, Jean-Pierre (1988). “Graphèmes et idéographie. Approche psycholinguistique de la notion de graphème”. In: *Pour une théorie de la langue écrite*. Ed. by Nina Catach. Paris: Éditions du CNRS, pp. 93–102.
- Joly, André (1982). “De la théorie du langage à l’analyse d’une langue. Remarques autour de la Grammaire de Condillac”. In: *Condillac et les problèmes du langage*. Ed. by J. Sgard. Genève, Paris: Slatkine, pp. 243–256.
- Mansour, Kamal (in this volume). “On the Origin of Arabic Script”.
- Massoudy, Hassan (1981). *Calligraphie arabe vivante*. Paris: Flammarion.
- Osborn, J.R. (2017). *Letters of Light. Arabic Script in Calligraphy, Print, and Digital Design*. Cambridge, MA, London: Harvard University Press.
- al-Qalqašandī, Abū l-‘Abbās Ḥamad ibn ‘Alī [أبو العباس أحمد بن علي بن أحمد] (1985). صبح الأعشى في صناعة الإنشاء [The Morning of the Dim-Sighted [Treatise] on the Art of Composition]. Cairo: الهيئة العامة للكتاب [al-Ḥay’at al-‘amma li-l-Kitāb].
- Robin, Christian (1991). “Les Écritures de l’Arabie avant l’Islam”. In: *Revue du Monde Musulman et de la Méditerranée* 61, pp. 127–137.
- Schoeler, Gregor (2002). *Écrire et transmettre dans les débuts de l’Islam*. Paris: Presses Universitaires de France.
- Vygotsky, Lev Semenovich [Выготский, Лев Семёнович] (1934). *Мышление и речь [Thought and Language]*. Moscow: Государственное Социально-Экономическое Издательство [Gosudarstvennoe Socialno-Ekonomicheskoye Izdatelstvo].
- (1935). “Проблема обучения и умственного развития в школьном возрасте [The Problem of Teaching and Mental Development at School Age]”. In: *Психологическая наука и образование [Psychological Science and Education]*, pp. 3–19.

Orthographies in Papua New Guinea through the Years

Ray Stegeman

Abstract. I would like to present the use of non-English graphemes by many other SIL-PNG colleagues over the years. SIL-PNG has been at work documenting little-known languages in Papua New Guinea for over 60 years. The motivation for using certain graphemes, diacritics and other writing strategies has changed through the years, particularly related to choices made in nearby languages, related languages and the choices available through the dominant, colonial language of English.

It is particularly interesting how in recent years, linguists and translators have moved away from using diacritics and other unique graphemes, especially with the advent of cell phone use, so underdifferentiation in the alphabet has become commonplace, and its relative acceptance and efficacy will be the focus of my presentation.

I have done a systematic analysis of graphemic choices made for most SIL projects in PNG for the past 60 years. Unusual, non-English graphemes are the focus of that research, and a questionnaire was sent to current SIL members asking about their motivations for using or not using certain unusual graphemes. I wish to compare which uncommon graphemes are chosen to represent which phonemes and gain insight into their efficacy as well as their general acceptance among the people who use the orthographies.

1. Introduction

SIL has been at work for over 60 years in Papua New Guinea, a land teeming with hundreds of languages. My colleagues and I have worked in over 300 of the over 800 languages that exist in this most linguistically diverse of countries/regions of the world.

As most everybody in the field of linguistics knows, Papua New Guinea is the most linguistically diverse nation on earth. The Ethnologue currently lists PNG as having 841 living languages. Of these, 164 are either “in trouble” or “dying”. Those are not particularly disparaging numbers, given that Wikipedia (<https://en.wikipedia.org/>

Ray Stegeman
SIL International
Dallas, TX, United States

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1. Fluxus Editions, Brest, 2019, p. 269–292. <https://doi.org/10.36824/2018-graf-steg>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

wiki/Endangered_language) reports that 50–90% of the world's languages will be gone by 2100. It seems that language identity and loyalty in PNG is still a very strong social force, and people here have a binding, common identity with their wantoks (people who share “one talk” with them) that overcomes many of the obvious forces of nation building and national identity through promotion of a national language or languages elsewhere. Western culture and psyche focus on “being on the same page” and “speaking the same language”. This is a strong cultural goal for us, particularly in the internet age, so that any linguistic individuality we may have falls to these preeminent pressures. In Papua New Guinea, not so much.

SIL International (formerly The Summer Institute of Linguistics) has been at work in PNG since 1956. By 2006, we had served in 337 language communities and we are currently serving in 187 of those same communities. Much linguistic work has been done on almost all of these languages (<http://www.pnglanguages.sil.org/resources>) including many grammars, phonologies and literacy materials. Along with those, there is a formidable amount of material on the design and reform of the orthographies that would best serve these many groups in reducing their languages to writing. You might imagine that when small communities pride themselves in preserving their own languages, they might also be interested in having their orthographies stand out as unique—different from nearby neighbors, if not necessarily different from the national languages of English, Tok Pisin and Hiri Motu. So, in surveying these orthographic materials, one finds a good number of strategies used for writing the myriad of languages in this country.

Many of our personnel have been and continue to be committed to seeing the indigenous languages of PNG written so that the speakers can feel a sense of pride at seeing their mother tongue on paper, in books, words in a dictionary, and so forth, and realize more completely the value of their language. Of course, this is not the only way to recognize the value and prestige of a language, but it is one very important way. And so, all those hundreds of languages must have orthographies.

Before being allocated to one of these language groups, my colleagues all had linguistic training, and we might admit that some of us were overzealous in trying to make plain all the esoteric phonological qualities of any one language in its orthography. While making it easier for an outsider to read and write in their language, this concept may actually make it more difficult for the local speakers to write and/or read their first languages—some of which are arguably among the most morphologically complex in the world. One colleague reports that in the language he studied, they can have more than 10,000 forms of any one verb (Menya, Whitehead, pc).

So, among these challenges arises the unique challenge of writing these languages—allowing them to learn from and share with each other

in ways that weren't available to them before—for parents to write a letter to their child, far away at a national high school; for someone to write to a brother working in the mines across the island. And, of course, with written Bible translation, people can check for themselves the Bible stories so often talked about. "Look, it's that story I heard about; it's written down right here, and I can read it for myself—and I understand it clearly!" Whether it be the Gospel of Mark or a community meeting notice at the local shop or checking up on what their children are learning at the local tokples preschool, they have a new-found power and prestige of seeing their language, not only talking it. They are not only preserving their language, they are expanding the use of their languages into new milieus.

As an introduction, I want to discuss some of the literature on the subject of orthographies in general and move toward more specific ideas coming out of PNG and SIL's work in PNG, since I have most access to our colleagues' work. After that, I'd like to present the data and some results I have gleaned from the data. I want to include some evidence from local testing of graphemes and especially subjective feedback from the SIL teams and the communities they worked with on this alphabet enterprise—the very people who are left to use the alphabet either given them, or the alphabet they choose, and more ideally, the orthography that represents the best of both efforts.

In following sections, I will present the use of non-English graphemes by many other SIL-PNG colleagues over the years. The motivation for using certain graphemes, diacritics and other writing strategies has changed through the years, particularly related to choices made in nearby languages, related languages and the choices available through the dominant, national language of English and trade language Tok Pisin. It is particularly interesting how in recent years, especially with the advent of cell phone use, underdifferentiation in the orthography has become more common.

I have done a systematic analysis of graphemic choices made in many of the SIL projects in PNG for the past 60 years. Use of non-English graphemes are the focus of that research, and a questionnaire was sent to current SIL members asking about their motivations for using or not using certain unusual graphemes and the graphemes they chose for sounds not described in the English alphabet. I wish to compare which uncommon graphemes are chosen to represent which phonemes and gain insight into their efficacy as well as their general acceptance among the language communities that use the orthographies.

2. Literature on Orthography Development

The idea that local language writing systems should resemble official and/or national languages is a well-established goal for any local orthography (Grenoble & Whaley, 2004:158 and others). For that reason, I won't investigate the possibility of languages in Papua New Guinea using any other kind of writing system, such as abjads or logographies, etc. It seems obvious they would not be helpful in allowing the language communities of PNG to bridge¹ between their local languages and the national languages of English and Tok Pisin. There has been some recent success in teaching the reading and writing in Uniskript (R. Peterson and D. Petterson 1994; also <http://uniskript.org/>), especially as a precursor to learning to read and write using a more regular alphabet. The Uniskript alphabet encourages its users to adapt the symbols to how a particular language community views the sounds they make, so although the "common" way of representing a t sound is with a triangle (showing the tip of the tongue touching the roof of the mouth), it can be adapted in different ways to suit the perceptions of sounds in each language community.²

Going back to the formative years of SIL, Pike (1947) wrote a book that included a chapter titled "The Formation of Practical Alphabets" which included linguistic and non-linguistic motivation for designing alphabets for local communities. He also addressed important topics such as dialects, knowledge transfer of first language reading and writing to languages of wider communication and advice against excessive use of diacritics, among other things.

A book by Smalley et al. (1964) called *Orthography Studies* addresses many topics including a chapter called *Practical limitations to a phonemic alphabet*. He suggests that many other sociolinguistic factors be considered when forming a writing system among a language community.

Gudschinsky's book, *A manual of literacy for preliterate peoples* (1973, has many helpful chapters on such literacy ideas as functional load, underdifferentiation, orthography testing and phonemic as well as morphophonemic representation. This book was adapted to language work in PNG and published by SIL in PNG. The sources from Pike and Gudschinsky both are very practical in their information on implementing orthography development and reform.

1. Bridging, here, is referring to the idea that mother-tongue orthography and literacy be designed as a reflection of the national/official language as much as possible, so that letters and spelling rules chosen for the mother-tongue help a student move more easily from mother tongue literacy to national/official language literacy and vice versa.

2. Unrelated to Uniskript, but on the same topic, one language community chose the tilde over their vowel letters to symbolize nasalization, and they call it *titi*—'wave,' and the symbol seems to them to represent a wave (Kala, Mitchell Michie, pc).

Simons (1977) suggests viable solutions for orthographies for multiple dialects, which is a highly salient topic in many areas in PNG. His multidialectal approach to orthography design can help closely related dialects to use the same orthography following seven principles that include social acceptability, minimal potential ambiguity, simplicity and phonemic contrast and neutralization between dialects. He also discusses four levels of psycholinguistic reality—phonetic, phonemic, morphophonemic and fast speech. He stresses that it is possible for dialects to have the same reality of a word in one of these realities, even when in other ones they would be different.

UNESCO has sponsored many articles and books on literacy and orthography development, one of which is *The manual for developing literacy and adult education programs in minority language communities* (Malone, 2004). It has a chapter on devising alphabets for previously unwritten languages, giving practical guidelines and helpful case studies from around the world.

Also from UNESCO is *Writing unwritten languages: A guide to the process* (Robinson and Gadellii, 2003), which has very practical help for grassroots level workers. It focuses on the stakeholders—those who would be using the orthography—and the issues that most often affect acceptability of the orthography. They also mention non-linguistic factors that can most often affect—positively or negatively—the acceptance of an alphabetic orthography. These include social, political and cultural considerations. Most of these are written in the form of questions, to get the language communities thinking for themselves, such as:

- Socially, what are the relationships between the language community and its own dialect areas? How do they view their language and possible orthography choices? What about the relationships between the language community and other languages within the greater area?
- Politically, what degree of autonomy does the language community have in making orthography decisions? What about colonizing effects on national as well as local literacy?
- Regarding preserving a cultural heritage, what does writing do to an exclusively oral culture? Who are the gatekeepers for preserving this newly written base of knowledge, when it has up to now been held by those who preserve it orally? (ibid., Section 3, p. 11–13)

In writing about language structure, Robinson and Gadellii speak about these important ideas for consideration:

- Representing distinctions in sounds to avoid confusion (minimize under-differentiation).
- The phonemic ideal that one symbol represent one sound, always and only.
- How does the grammatical structure of the language influence how it is written? (e.g., word breaks, elision, verb morphology, etc.)
- Can the structures of related languages give ideas for developing the orthography? (ibid., Section 3)

Cahill (2014) mentions that in addition to sound scientific study and community involvement, effective teaching materials and practices are crucial to having a lively, sustainable literate public. He also talks about promoting the following benefits to a language community that is seeking to own and utilize mother-tongue literacy:

Tools to deal with the larger world which is unavoidably coming (to all language communities) in:

- interacting with computers and the internet;
 - thwarting attempted land grabs and other legal tangles;
 - more math awareness, to deal with others, especially regarding money;
 - more access to beneficial materials;
 - gaining knowledge about health (water, AIDS, nutrition, etc.);
 - knowing the contents of government and NGO documents (e.g., UN Declaration of Human Rights);
 - not losing almost-forgotten folk tales and other local lore;
 - preservation of a community's cultural resources;
 - strengthening one's cultural identity and having a higher view of one's own language;
 - strengthening the language (Malone (2004) lists "materials for language education and literacy" as one of the nine factors that most affect language vitality);
 - for women especially, and sometimes whole language groups, increased self-esteem;
 - letter-writing can be more private than conversations;
 - an aid when traveling (e.g., following directions, finding destinations, etc.).
- (Cahill, 2014, Appendix, p. 6)

Other literature was also consulted while writing this paper; please see the bibliography for details.

3. PNG Milieu

There are over 800 living languages in PNG—a nation of less than 8 million people, occupying an area less than the size of France. There are obviously many challenges to nation building and communication among this ethnically and linguistically diverse people. Still, many of the language communities have been and continue to stay connected by trading and using common trading languages like Tok Pisin and Hiri Motu. In the more recent past, with roads connecting communities and better infrastructure, more and more people are moving to the bigger cities for economic and educational opportunities. These people, while desiring to keep their cultural and linguistic heritage, are also realizing the opportunities for advancement by way of modernization and living in multi-linguistic and multi-cultural communities.

Many of the languages are related to many others, so even though one can walk down the road just 5–10 km and likely find another group

of people speaking a different language, many times, the languages are related. Many of the same cultural ideas are preserved in multilingual communities (e.g., trade, the cultural custom for men to marry outside their clans, etc.) These ideas feed the concept that many people are multilingual, depending on the environment and the audience. And so, many Papua New Guineans have a knack for learning languages, appreciating sound and grammar patterns that those of us with fewer multilingual opportunities could easily miss or overlook.

The official languages of Papua New Guinea are English, Tok Pisin and Hiri Motu. Tok Pisin is an English-based pidgin/creole and used largely in the north and highlands areas, while Hiri Motu is an adaptation of Motu, an Austronesian language, developed through trade and by colonial efforts mostly in the south of the country. All three use similar orthographies and spelling rules, and these are taught in schools, so any literate person working on indigenous alphabet development will be strongly influenced by these traditions when it comes to creating an alphabet in one's own language for the first time.

The two biggest language families in Papua New Guinea are Trans New Guinea (483, *Ethnologue.com*)³ and Austronesian (1,256—not all in PNG, *Ethnologue.com*). Austronesian languages are mainly found along the coast and in the outlying islands of PNG, while Trans New Guinea languages are largely found in the highlands. There are some exceptions to this, and one can find largely Austronesian languages displaying some obvious Trans New Guinea features (similar pronouns and counting systems, word order, grammatical features, etc.) and vice versa. This might seem natural in a place like PNG, since over time there has been so much language contact among the different communities. It would be and indeed is, hard to tease apart the indigenous features from the borrowed features of any individual language.

One feature of interest for this study is that the number of graphemes is largely related to the number of phonemes needing representation in the orthography. While Austronesian languages are said to be less phonologically complex than Trans New Guinea languages (<https://www.britannica.com/topic/Austronesian-languages/Structural-characteristics-of-Austronesian-languages>), I have found phoneme numbers to be widely divergent and seemingly not related to the family of any one language. One famous PNG language, Rotokas (N Bougainville language family), is known for having (arguably) the fewest phonemes of any language on earth—11. The Trans New Guinea language of Melpa has 26 phonemes. Alekano, also a TNG language, has only 16 phonemes while Sudest, an Austronesian language, has 40 phonemes. These are offered

3. Trans New Guinea is actually a grouping of many language families that are known not to be Austronesian.

as a small sample that confirms languages from either major language family can have a big difference in the numbers of phonemes.

One objective of this study was to look at how many non-English letters and diacritics are needed/used in a local orthography, so the number of phonemes, while far from the only consideration, is a significant consideration—possibly the best starting point for engaging the local communities in making alphabet choices.

4. SIL's Work in Papua New Guinea

SIL International began work in PNG as the Summer Institute of Linguistics in 1956. Following what Pike (1947) and others promoted, many of our SIL teams worked hard to discover the basic building blocks of each language they encountered—sounds, phonemes, morphemes, morphophonemic sound changes, and other linguistic concerns that are directly related to any orthography developed for the languages. Most of their work is documented and can be found on www.pnglanguages.sil.org.

While many of the teams were careful in their linguistic studies, and in accordance with their linguistic training, they also consulted members of the language communities in which they worked, eliciting feedback from local leaders as to the best ideas for alphabet choices. In the questionnaire I gave to 39 current SIL colleagues, 37 of them said they had ongoing consultations with local leaders—often school teachers and/or school administrators, respected village and church leaders—in making initial and ongoing decisions about alphabet letter choices and spelling rules. They worked together to address linguistic concerns as well as many sociological concerns with respect to the orthography. Many SIL colleagues spent many long years doing literacy and translation work among the people, living in their villages for months at a time. They have documentation of testing different alphabet choices made by them, in conjunction with the greater language community as to how they “liked” what they saw and/or were successful in learning to read and write. Many of the SIL teams were following the extensive testing procedures covered in Gudschinsky’s book (1973, ch. 13) along with other resources.

One of the issues that seemed to come up most frequently was how similar or different one’s language “looks” on paper compared with neighboring languages. Some language communities desired to follow what they saw being used around them and sought to use letters and diacritics similar to nearby languages. Other communities sought to “show off” the uniqueness of their languages by doing things differently from those nearby. So, as a non-linguistic factor, one of these two opposing

preferences could be mainly responsible for how particular items in a single orthography come to be used.

More recently, with the advent of large-scale cell phone use in PNG, it has become a concern of many language communities how relatively easy or difficult it is to text in their mother tongues, without the aid of apps that need to be downloaded and often set up (InKey, Keyman, etc.) This makes it a priority to use letters and diacritics available on simpler, lower-priced cell phones, to allow those who wish to text in tokples ('talk place'—the local language). This was the topic addressed most in the Questionnaire section of my research for this paper.

5. Grapheme Choices

In looking at SIL phonologies and orthographies of PNG languages through the years, I was interested in charting strategies that teams used to describe sounds that required non-English graphemes. These included diacritics, digraphs and trigraphs (and one tetragraph) that are not used in English, which I group together to call multigraphs from here on. I was also interested in the use of English letters in the orthography used for sounds other than regular English sounds, as in for the phoneme /β/, or <c> for the glottal stop. I also looked at the strategies of overdifferentiation and underdifferentiation, to see if there were any trends developing through the many years SIL has been at work among PNG languages.

I had access to three different sources of this orthography data, based on the relative dates of the information. The oldest material comes from charts of phonemes mapped with graphemes of language projects from before 1990. The second group of data comes from OPDs (Orthographic and Phonology Data) which are on file in the Linguistics Office at Ukarumpa. Many of these OPDs have been written and/or updated up to 2010. The third group of data available to me came from answers to a questionnaire given to current language teams and the current graphemic data they provided. The time frames overlap from the standpoint of when the orthographies were developed, but the endpoints are fixed and exclusive, so that orthography strategies from the first group (before 1990 would not have been concerned with, say, relative ease of texting in the mother tongue, whereas data from answers to the questionnaire (that is, data from a currently active language project) would be concerned about strategies for employing graphemes in direct relation to that issue or others. So, I believe the three groups of data correspond with different philosophies of literacy, reading and writing, and communication tools available to the local speakers/readers/writers in separate time frames.

Along with their graphemic data, those who completed the questionnaire gave answers to questions regarding who was involved in the language program and the grapheme choices as well as changes made to their graphemes throughout the life of the program and the reasons they felt they needed to make changes. A list of the questions from the questionnaire is together with the answers in section 8 below.

Regarding the grapheme data, I counted the number of instances of each strategy (“strategy” here refers to things like using a tilde to show nasalization, or the digraph <ng> for the velar nasal phoneme /ŋ/) and kept track of them on a chart. Then I put these individual strategies into major groups (“major groups” here refers to diacritics, multigraphs and letters not used elsewhere). Some interesting results emerged as I put the data into six groups and compared their numbers in relation to the 3 historical eras. I needed to give each section a weighted average, since the number of languages researched was different from those in the other two groups.

One caveat: The languages in this survey are from a number of language families (charted below in section 6), and the percentages from the different language families are not constant in the 3 time-related sections (the middle three columns). It is possible that the increased number of Austronesian languages from the newer data (Questionnaire) section of research was the main reason for, say, the increase in diacritic use from the older data, which more heavily favors Trans New Guinea languages. This skewing factor could be eliminated in any future study of a similar nature.

6. Data Collection

Table 1 shows the number of languages in my research and in what province they are mainly located.

Table 2 shows the language families to which the researched languages belong. Again, Trans New Guinea is a convenient designation for several large language families traditionally recognized as non-Austronesian.

And finally, Table 3 shows the actual research data. The numbers in columns 4–7 represent the raw data (RD) of the research and the second number in each cell represents the weighted mean (WM), based on the different number of languages in each section divided by the total number of languages, so the second number in each cell are the actual numbers for comparison across columns.

TABLE 1. Languages from research and in what province they are mostly spoken

PNG Province names	older data ←		→ newer data	
	Languages from before 1990	Languages from before 2010	Language teams responding to the Ques- tionnaire (current projects)	Total languages from provinces
Morobe	5	9	9	23
East Sepik	9	7	3	19
Madang	2	15	2	19
Milne Bay	5	9	4	18
Gulf	7	6	2	15
Western	5	5	4	14
Autonomous Region of Bougainville	7	2	3	12
East New Britain	8	1	3	12
Eastern Highlands	7	4	1	12
Central	10	1	1	12
Sandaun	4	7		11
New Ireland	7	1	3	11
Manus	6	2	2	10
Oro	6	2	1	9
Simbu	6	1	1	8
Southern Highlands	5	1		6
West New Britain	3			3
Western Highlands	3			3
Enga	2			2
Hela			1	1
Jiwaka	1			1
Total	108	73	40	221

7. Data Results

7.1. Increase in Use of Diacritics

Moving from the past to the present (moving across Chart 3, from left to right), the data show an increasing use of diacritics among the orthogra-

TABLE 2. The number of researched languages and the language families to which they belong

PNG language families represented	older data ←		→ newer data		Total in language families
	Languages from before 1990	Languages from before 2010	Language teams responding to the Questionnaire (current projects)		
Trans New Guinea	53	37	11		101
Austronesian	35	15	21		71
Sepik	6	4	0		10
Torricelli	1	4	3		8
South-Central Papuan	0	3	3		6
Isolate	3	1	0		4
Ramu-Lower Sepik	1	3	0		4
South Bougainville	2	1	0		3
Border	1	1	0		2
Eastern Trans-Fly	1	0	1		2
Senagi	1	1	0		2
Yele-Western New Britain	0	2	0		2
Arai	1	0	0		1
East New Britain, Baining	0	0	1		1
East New Britain, Taulil	1	0	0		1
Fas	1	0	0		1
North Bougainville	1	0	0		1
Skou	0	1	0		1
Total	108	73	40		221
Divide raw data below by this number to get a weighted mean	.489	.330	.181		1.000

TABLE 3. Comparing use of each orthography strategy across the 3 time periods of research

Strategy used	Examples	older data ←		→ newer data		Change in use of each strategy
		Languages from before 1990	Languages from before 2010	Language teams responding to the Questionnaire (current projects)		
		RD/WM	RD/WM	RD/WM		
1. diacritic	ë, ã, ú	78/160	65/197	54/298		increase in use
2. multigraph	th, mp, ndr	264/540	263/797	153/845		increase in use
3. underdifferentiation (including phonemes not written)	<e> for both /e/ and /ə/	40/81.8	35/106	33/182		significant increase in use
4. overdifferentiation	 and <mb> for /b/	114/233	89/270	32/177		eventual decrease in use
5. English letter not used elsewhere	c, q, x	125/256	72/218	33/182		decrease in use
6. non-English letter	', ʔ, ŋ	36/73.6 (6 ŋ, 17%)	31/93.9 (13 ŋ, 42%)	8/44.2 (5 ŋ, 62%)		eventual decrease in use (but increase in use of ŋ)

phies SIL language teams employ. The weighted mean number (WM, the second number in each cell) increases. (Row 1 has the numbers 160, 197, 298, from left to right.) Some examples of diacritic use include certain marking on vowels to show nasalization, so that the languages have a set of non-nasal vowel letters, say <a, e, i, o, u> along with a nasalized set, say <ã, ë, î, ò, û>, or possibly a complementary set of long vowels

marked with dieresis, say <ä, ë, ï, ö, ü>. Of course, if a language team decide to represent nasality or length with diacritics, this adds significantly to the overall number of diacritics used—5 or more, as opposed to using a single diacritic to show, say, a dental t phoneme different from alveolar t. I chose to document the marking of nasalization and length (the two most common need for graphemic adjustment) by the diacritic used on each letter, as opposed to documenting use of only one diacritic. This is because some languages documented only some vowels as having nasalized counterparts and not the whole set of vowels.

Use of diacritics in orthographies employed by SIL-PNG language projects were mostly used in the vowel systems, either showing nasalization, as above, or a similar place of articulation on the vowel chart. An example of this would be <u> for the close, back, rounded vowel /u/ and <ü> (with a dieresis) for the close, middle vowel /i/. Gizrra does this, along with using an acute mark on the o <ó> for the schwa phoneme /ə/ while also having the <o> letter for the middle, back rounded vowel /o/ phoneme. Thus, for seven vowel phonemes, they use the regular 5 vowel letters of English and two of the same vowel letters with two distinct diacritics.⁴

7.2. Increase in Use of Multigraphs

There was also an increase in the number of multigraphs used in the alphabets chosen by SIL-PNG language projects in more recent years. Many PNG languages have phonemic systems that involve prenasalization /^mb, ⁿd, ^ŋg/ and labialization /p^w, t^w, k^w/ of plosive phonemes. These can have some phonemic alterations, such as word-initial, word-medial or word-final forms. This can mean that it is not phonemically critical to show the prenasalization or labialization in the alphabet as a matter of linguistic description, but it may be preferred by the language communities. Often, this is the case, due to the influence of English as an official language, and the letters they see when they read English words like *combine*, *condition*, *rwin*, *quick*, etc. and they see both sounds represented, they feel like their languages should be written the same way, with both letters used, even though the phonemic reality for the two languages can be quite different. Some PNG languages have both prenasalised *and* labialized consonants /^mb^w, ⁿd^w, ^ŋg^w/, which are sometimes represented as trigraphs, with one tetragraph used in one language for the prenasalized, labialized velar plosive, /^ŋg^w/, spelled

4. It is interesting to note that this team felt it was easier to recognize different diacritics for the extra two vowel phonemes, rather than using the same diacritic. Both the letter (o and u) and the diacritic show the difference, to emphasize recognition of the difference in reading more quickly (vanBodegraven, pc).

<nggw> (Khehek, Manus). As with nasalization on vowels, representing these phoneme series as multigraphs can have a multiplying effect on the number of counted strategies used in orthography design. It also tends to make certain words (more words in some languages than in others) unduly long, and the length is multiplied since the condition of prenasalization and/or labialization usually occurs over the whole range of plosives, and not just a single phoneme.

Use of digraphs is also a common strategy in vowel phonemes. Double vowels such as <aa, ee, ii, oo, uu> are often employed in orthographies to identify length. Use of double letters for vowel length, and digraph use in general, can lead to disproportionately long words, depending on the actual language and how many syllables have nasalization or length (or both) on the vowels. In fact, digraphs representing vowel length is a common strategy among SIL-PNG language projects.

Although it isn't phonemically necessary, language communities in a situation where English is the national language feel compelled to use nasal letters together with the prenasalized consonant phonemes. This is a valid consideration in relation to the official languages of English and Tok Pisin, especially when the concept of bridging between languages is a serious concern.⁵ Local speakers have learned about spelling rules in English, and they have learned to read English from being taught in school. When they hear the same sounds in their language, they tend to want to use the same letters and spelling rules that they know from another language, particularly the prestigious official language. This can be alright in some cases, and it can be more helpful in a multi-lingual environment like PNG, where people move easily from language to language based on the social situation, but such uninformed orthographic transfer can ignore the unique phonemic and morphophonemic tendencies of any one language, which could help in more quickly and completely acquiring fluency in reading and writing one's mother tongue. Languages, after all, are so very different from each other; it seems natural (to an outsider/linguist) that the orthographies representing these languages would also be very different from each other. It is probably best from a bridging perspective to utilize some of the same letters for the same phonemes (*not* sounds) while also showcasing other unique phonemic qualities of a language by making some unique grapheme choices, which could include the writing of prenasalization.

5. Bridging, here, is referring to the idea that mother-tongue orthography and literacy be designed as a reflection of the national/official language as much as possible, so that letters and spelling rules chosen for the mother-tongue help a student move more easily from mother tongue literacy to national/official language literacy and vice versa.

7.3. Increase in Use of Underdifferentiation

This seems like a surprising result, considering the previously mentioned two trends. One might think that using more diacritics and multigraphs would correspond with (adequate) differentiation or even overdifferentiation, as I have been talking about above.

The increase in use of underdifferentiation could be the result of the increased use of technology having a direct effect on language and literacy development, and the felt need for communicating in one's mother tongue using different electronic devices, including cell phone use. Cell phone use has skyrocketed in PNG in the recent past, and it is a felt need, at least in some language communities, to use one's mother tongue in calling and texting each other. While the phonology of a particular language may be complex enough to need many diacritics and/or multigraphs for a more phonemic representation in the alphabet, it may be even more desirable by the community to reduce the number of "untextable" letters in the alphabet, to make it easier to communicate with each other by using today's technology. This is mentioned in a few of the questionnaire responses, as can be seen in section 8.4.

This underdifferentiation has a trade-off in that while it is easier to text/write, it is often much more difficult to read. The onus of communication falls to the reader in deciphering a message that could have more than one meaning based on the lack of sufficient letters for the meaningful sounds of a language. Many of the questionnaire respondents mentioned that speakers usually text without the diacritics in the official orthography, and for a few diacritics (depending on how many and how often they are left out) they can make themselves easily understood. This would obviously have a limit, so that by leaving off 5 or 8 or 12 special characters or digraphs, one's texting would certainly become more of a deciphering challenge than actual effective communication.

7.4. Decrease in Use of Overdifferentiation

Both increasing underdifferentiation and decreasing overdifferentiation could be tied to a growing interest in using modern technology in one's mother tongue without the need for special apps. More and more Papua New Guineans have access to computers and smart phones, but they don't necessarily have the ability or knowledge to adapt them for mother tongue language use. To have an orthography that is simple enough to use a regular cell phone to communicate in one's mother tongue seems to be more of an interest than before.

Some teams also mentioned the desire for shorter words. A lot of PNG languages can have complex morphology, especially on the verb, and this can make words unwieldy in their length. Together with multi-

graphs, written words become difficult to read. One way to counteract this problem is to use fewer multigraphs, which might be preferred for other reasons (like bridging) but would decrease the use of overdifferentiation.

7.5. Decrease in Use of English Letters Not Used Elsewhere

Over the decades, there has been a noticeable decrease in the use of English letters not used elsewhere in the mother tongue grapheme inventory for SIL-PNG language teams. These letters often include v, z, q, c, x, w, and/or y. In the past, using one of these letters was often the strategy for indicating the glottal stop phoneme, /ʔ/, for which some languages use <q> or <c>. Some languages use <x> for uvular phonemes like /x/ or /ɣ/. This seems like a good strategy to use, especially based on current texting concerns. These letters are immediately available on regular computer keyboards and texting devices, and they don't require special apps or set-up. But the data in this survey shows they are not used as often as they were in the past. This could be due to the bridging concerns mentioned earlier, where letters used in one's mother tongue are expected to reflect the alphabet and sound patterns used in the official language(s). So, for example, using a <c> for the glottal stop doesn't "feel" natural, when one has a strong association that the c letter should/must represent the [k] sound as in <cat> or [kæt] and not the glottal sound. Of course, the major difference is that the glottal stop is not a phonemic sound in English, while it is a meaningful sound in many PNG languages, and necessary to include in the orthography for that reason.

7.6. Decrease in Use of Non-English Letters

The use of ɲ as a grapheme has increased over time, which contrasts with the overall decrease in using other non-English letters. These include the apostrophe or question mark (both mostly for the glottal stop), certain IPA graphemes (usually for similar sounds/phonemes) including letters like <ɬ, æ, ə> etc. The decreasing use of letters like these in PNG orthographies is perhaps to be expected, again considering the spread of technology and the texting phenomenon. These characters for use in an alphabet are not standard on computer keyboards or phone touchpads. Some special, non-English letters are found on smartphones by pressing and holding certain letters, which reveals a choice of alternate characters, but this feature is only available on higher-end smart phones and often only with the special characters used in European languages; that is, mostly English letters with certain diacritics and no non-English letters like above.

It may seem contradictory that with an increase in the availability of technology, we would have more choices in the alphabet letters available to us. But in fact, it seems like we are still bound to using only a standard computer keyboard (like we were bound to using typewriter keyboards before computers) and bound to using only a cell phone keypad with the only options being European diacritics favored by the phone makers or stakeholders other than those people with emerging writing systems for their languages. While Unicode has increased the availability of a vast array of characters for use, the vast majority of them are still not available to the average language speaker who would like to text/write in his/her first language, but who doesn't have or know about the options for using all those characters on simple technology.

7.7. Overall Impressions

Based on the questionnaire responses (newer data) versus older orthographic data, there seem to be forces at work causing language projects to use strategies for crafting orthographies that are less purely based on phonemics, more based on mirroring official languages, and more a reflection of the strong felt need for simpler orthographies for use on technology currently available to language communities. These sociolinguistic forces are made evident in the increasing use of digraphs, the increasing use of underdifferentiation and the decreasing use of overdifferentiation.

8. Questionnaire Feedback

8.1. I asked the currently active SIL teams to share with me some general ideas about the specific situation they found among the local speakers of the languages with whom they work. The idea they mentioned most was the challenges they face trying to develop an alphabet for multiple dialects. One team mentioned the people all had strong dialect loyalty, which meant it was hard for the speakers of various dialects to utilize a common alphabet and spelling rules that didn't reflect their particular pronunciations.

A couple teams mentioned that they had no significant dialect challenges and were able to realize a unified orthography across minor dialect boundaries.

Another challenge to teams was that of having had multiple SIL teams at work in the same language communities through the years. There were also instances of other mission agencies working in the same area previously, in particular, German missionaries who made alphabet choices based on German sound-symbol correspondences (e.g., <ch> for

/x/), which make it difficult for transferring reading skills from mother tongue to the official language now that it is English. Another example was the influence of certain Fijian missionaries who chose <g> to represent the velar nasal /ŋ/. Some choices made by earlier teams are often difficult to overturn, especially as the older generation owns and appreciates the earlier choices, but the younger generations would like to see something based more on current realities.

8.2. I also asked current teams about the stakeholders that were involved in the alphabet-making enterprise. Who gave input into the process? How were they chosen?

Those who became involved (other than the SIL team themselves) were mainly school teachers, local speakers of the language (through informal, occasional meetings), and local church and other community leaders. Many of the SIL teams formed either language committees or translation committees who were responsible for making orthography decisions, often meeting on a regular basis and conducting surveys or tests related to orthographic choices and/or changes.

The SIL teams have many tools at their disposal to help make or force decisions about the orthography. Some of the more common tools mentioned as being instrumental in the process are: Alphabet Development Workshops (Nukna, Blafe, Middle Kodut and others), developing tokples prep school literacy materials (Nek, Seimat, Nehan and others), orthography testing methods (Edolo, Mato, Seimat, Iyo and others), writers' workshops (Misima, Solos, Lote, Arop-Lokep, Kanja, Nehan) and distributing copies of the trial orthography along with locally authored stories using the orthography and then getting feedback from fellow speakers about their alphabet preferences.

8.3. The next question had to do with orthographic strategies they tried early on and decided to change. There were not many common answers in this section, as a testimony to the linguistic diversity and language communities' unique preferences. The idea of overdifferentiation was mentioned most often, particularly for prenasalization, pre-clusivized nasals and nasalized vowels. It was determined in most of these cases that readers benefit from "seeing" the overdifferentiation, likely because of what they had learned from literacy in the official language in their school education.

There were some decisions made in one language community for which another community made a contrary decision. The two strategies mentioned in the answers to the questionnaire were marking nasalization on vowels and the orthograph chosen for the velar nasal. One team hadn't marked nasalization on their vowels initially but finally decided they preferred marking them. The reason given was that the speakers preferred seeing/reading the difference. Another team, while at first marking nasalization changed so as not to mark it. They made

this decision as a concession to writers, to make their job easier, but it no doubt gave readers more work. It was mentioned elsewhere in the questionnaire replies that writers often leave out certain letters—often those representing overdifferentiation—and that the readers are often able to read the materials anyway (Gizrra, among others). This was mentioned most often when talking about texting. It is also common among Westerners (at least in my home country, USA, and for many PNG people, messaging in Tok Pisin) texting and Facebooking to take shortcuts in their writing. Maybe these language communities in PNG have become familiar with others' texting and posting habits in other languages, and they learn to be brief (including underdifferentiation) based on other people doing the same in English. It could be a unique way of expressing oneself, even if others must work harder at deciphering the content.

A lot of angst was expressed when dealing with a language that is found to have more than 5 vowel phonemes. In some of these cases, it was decided that the vowel phonemes should be underdifferentiated. Dadibi has phonemic nasalization on all 5 vowels, but they choose not to write the nasalization. In other language communities, digraphs of vowel combinations or of the same vowel were used for a vowel phoneme of a similar quality. Ambulas, for example, uses <a> for the vowel phoneme /ɐ/ and <aa> for the /ɑ/, while Gapapaiwa uses <i> for /i/ and <ii> for /i/. In an effort to keep words as short as possible, it is a common practice when using double letters to use them for the less common phoneme.

It was mentioned in one team that a previous orthographic influence was from Fiji, and they found it necessary to move from orthographs common in Fijian languages to others, since that social influence was no longer in effect here in PNG (Mussau-Emira). It was determined that it is more important to have letters that help speakers bridge from their mother tongue to the official language of English, so the letter <g> for the phoneme /ŋ/ and the letter <q> for the phoneme /ɣ/ were not helpful in today's linguistic climate.

8.4. The next question in the questionnaire for current language teams was about technology and texting using the orthographies they had supervised. The most common answer was that there are no special characters or diacritics in the orthography, so texting using one's mother tongue was no special challenge in that regard. The next most common response was that the speakers were experimenting with texting by not using any of the diacritics and "getting by" with the lack of differentiation. This includes using <ng> for texting the velar nasal phoneme, even when their orthograph for this phoneme is <ŋ>. Whether or not the local speakers were feeling successful about this enterprise was not mentioned. One community (Gizrra) that had only two diacritics (an acute and a diaeresis) were successfully texting without these diacritics, and

they were making themselves understood. I suppose the relative number of diacritics and special characters in the official orthography would determine the relative success of this endeavor in any language.

One enterprising community was using numbers to reduce the effort texting takes, especially when the words are long due to reduplication. For example, for the word <waiwaisana>, they were texting <wai2sana>.⁶

8.5. The final question in the questionnaire for current language projects was about the major influences on the orthography, whether mostly phonemic or more a result of language community input. It seemed that of all the stated responses in current teams, the influence was about equal between the linguistic/phonemic influence and community input. This pairs well with the literature on the issue, that often states that linguistic and non-linguistic forces are at work in language development in general. Of course, the orthography has everything to do with how a language looks to its readers and their perception of how their language looks to the outside world—through their writing system. A language community must feel confident that the orthography they are using to showcase their language is adequate and practical but also a personal expression of themselves through language. It's not just a string of sounds—it's my language.

Questionnaire responses also talked about wanting to have reading easier and/or the teaching of reading and writing to be easier. This seemed to be important among language communities that also wanted an easier bridge to reading and writing in the official language. This desire seems to point to a need for more of a phonemic influence on the orthography, which would be a pull in the other direction from many of the other influences mentioned in this paper.

8.1. Conclusion

Grenoble and Whaley (2004, p. 158) list five recommendations at the end of their chapter on orthographies that I think are representative of the

6. *Note by the Editor.* The convention of using a <2> to reduplicate the graphemes preceding it has been attested in Latin-script Malay: according to Haji Omar (1989, §10), "There are three types of reduplication in Malay: the reduplication of the first syllable of the root, the reduplication of the stem of a complex word, and the reduplication of the whole word, be it a simple or complex word. In the old spelling systems both in Malaysia and Indonesia, the first type of reduplication was spelt in toto, but the character <2> was used to indicate the reduplication of the second and third types. In the reduplication of the whole word, the character <2> was placed at the end of the word, for example, <rumah2> was read as *rumah-rumah* 'houses,' <makan2> as *makan-makan* 'to while away the time eating'."

SIL experience in PNG. The first item is that orthographies devised to be used in a language revitalization project should be focused primarily on utilizing an *alphabetic* system. This goes without saying in PNG, where bridging concerns are primary in the thoughts of language describers/documenters, the government and the language communities themselves. This was obvious in the responses to my questionnaire. The orthography of a language in this situation will be used by many semi-literate people, so it needs to be instructive, teaching and reinforcing a speaker's knowledge of the sounds of one's language, and s/he will pick that up most easily from an alphabetic orthography. This point is well-accepted in PNG, so it needs no further discussion here. All our SIL-PNG projects have adopted this stance.

The second characteristic of a successful orthography they mention is *learnability*, where the orthography helps and encourages the learner in any way possible. Motivation can easily be discouraged if the enterprise of reading seems too difficult. Languages in PNG being revitalized through the use of a new orthography do not have the advantage of a well-established national or official language, where learning to read and write offers its own rewards of being in touch with the greater world through books, movies, internet, etc. Local language learners must be encouraged through any means possible to learn to read and write their own mother tongue, and the orthography design can aid in this process. Our SIL-PNG teams have shown this concern in their answers to my questionnaire by showing a real desire for successful local level literacy, working together with the language communities to make sure alphabet choices reflect the community's wishes, including different and often competing ideologies.

The third point they make is that orthographies should be *phonemic* as much as possible. The meaningful sounds of the language should be evident in the orthography, particularly those items with a high functional load. Depending on the language, application of this principle could be in tension with Grenoble & Whaley's second point above. A language in which there are relatively more phonemes doesn't allow for a simple, easily learned alphabet and spelling rules. But a phonemic understanding and basis is a good starting point, and most of the simpler issues can be easily addressed with this point in mind from the beginning.

Point four speaks of *transparency* in that "spelling conventions should coincide with those of the language of wider communication wherever possible" (p. 159). This is the same "bridging" concern mentioned by SIL-PNG teams in responses to the questionnaire. Any benefit to having a unique system of reading and writing (in competition with the national language) is offset by the limitations on how it helps or hinders the literate person. If the orthography is transparent, the language learners can become literate in both their own language and the official language(s). The skill of reading and writing can transfer more easily

to another language that shares a mostly common alphabet. Many SIL-PNG teams have made this a priority in their orthography designs. They are using many of the digraphs we use in English, such as <ng> for the velar nasal.

Their last point refers to the *acceptability* of the orthography. A writing system is only successful if it is acceptable to the language communities that are motivated to learn it and use it. Our SIL-PNG teams showed this to be a constant concern in their orthographies and in their literacy programs. We need to continue consulting with the language communities, so that the factors that concern them get incorporated into the orthography and allow them to realize the language revitalization they desire.

References

- Cahill, Michael (2014). "Non-linguistic Factors in Orthographies". In: *Publications in Language Use and Education*. Vol. 6: *Developing Orthographies for Unwritten Languages*. Ed. by Michael Cahill and Keren Rice. Dallas, TX: SIL International, pp. 9–25.
- Cahill, Michael and Keren Rice (2014). *Developing Orthographies for Unwritten Languages*. Dallas, TX: SIL International.
- Clifton, John M., ed. (1987). *Data Papers on Papua New Guinea Languages*. Vol. 33: *Studies in Melanesian Orthographies*. Ukarumpa: SIL.
- Grenoble, Lenore A. and Lindsay J. Whaley (2004). "Orthography". In: *Saving Languages: An Introduction to Language Revitalization*. Cambridge: Cambridge University Press.
- Gudschinsky, Sarah C. (1973). *A Manual of Literacy for Preliterate Peoples*. Ukarumpa: SIL.
- Haji Omar, Asmah (1989). "The Malay Spelling Reform". In: *Journal of Simplified Spelling Society* 11, pp. 9–13.
- Karan, Elke (2006). "Writing System Development and Reform: A Process". Master's thesis. Univ. of North Dakota.
- (2014). *The ABD of Orthography Testing: Practical Guidelines*. Vol. 54. Dallas, TX: SIL.
- Larsen, Robert E. (1977). *Multidialectal Orthographic and Lexical Adjustments for Orokaiva*. Vol. 21. Ukarumpa: SIL, pp. 343–348.
- Litteral, R. and Susan Malone (1991). *The Sounds of Your Language*. Port Moresby: Department of Education.
- Malone, Susan (2004). *Manual for Developing Literacy and Adult Education Programmes in Minority Language Communities*. Bangkok: UNESCO.
- Petterson, Robbie and Debby Petterson (1994). "Failures and Successes in Literacy in Gulf Province Schools". In: *Conference of the Linguistic Society of Papua New Guinea, Madang*.

- Pike, Kenneth L. (1947). *Phonemics: A Technique for Reducing Languages to Writing*. Vol. 3. Ann Arbor: University of Michigan Publications.
- Roberts, John R. (2002). *Orthography Reform in Amele*. Ukarumpa: SIL.
- Robinson, Clinton and Karl Gadelii (2003). *Writing Unwritten Languages: A Guide to the Process*. Paris: UNESCO. <https://pdfs.semanticscholar.org/bda5/26059b80037af03b0eaf4fec84ab696bf114.pdf>.
- Sarvasy, Hannah and Diana Forker, eds. (2018). *Word Hunters*. Amsterdam: John Benjamins.
- Schreyer, Christine (2017). "Reflections on the Kala Biṅatuwā, a Three-Year-Old Alphabet from Papua New Guinea". In: *Creating Orthographies for Endangered Languages*. Ed. by Mari C. Jones and Damien Mooney. Cambridge: Cambridge University Press.
- "SIL Papua New Guinea. Annual Report" (2017).
- Simons, Gary (1977). *Principles of Multidialectal Orthography Design*. Vol. 21. Ukarumpa: SIL, pp. 325–342.
- Smalley, William A. et al. (1964). *Orthography Studies: Articles on New Writing Systems*. London: United Bible Societies.
- Snyder, David (1994). "Orthographic Symbols in Papua New Guinea Languages". In: *Conference of the Linguistic Society of Papua New Guinea, Madang*.
- Spilioti, Tereza (2009). "Graphemic Representation of Text Messaging: Alphabet Choice and Code Switches in Greek SMS". In: *Pragmatics* 19, pp. 393–412.
- Tomokiyo, Laura Mayfield (2018). "Orthography Development". Presentation <https://slideplayer.com/slide/13082277/>.
- Venezky, Richard L. (2003). "In Search of the Perfect Orthography". In: *Written Language and Literacy* 7, pp. 139–163.
- Whitehead, Carl R. (2004). "A Reference Grammar of Menya, an Angan Language of Papua New Guinea". PhD thesis. University of Mani-toba.

Marking Tone with Punctuation: Orthography Experimentation and Reform in Eastern Dan (Côte d’Ivoire)

David Roberts, Dana Basnight-Brown & Valentin Vydrin

Abstract. Eastern Dan has five level tones, six contours and many monosyllabic words, resulting in an extraordinarily heavy functional load of tone. This led those first involved in orthography development to create a novel system for marking tone that uses punctuation symbols in word-initial and word-final position. This orthography also has considerable segmental over-representation and makes extensive use of umlauts to symbolize vowels. In a quantitative classroom experiment, we tested it against Valentin Vydrin’s recent proposal for radical reform that advocates superscript diacritics for marking tone, biunique correspondence for consonants and vowels, and special characters in place of umlauts. Sixty-eight participants with no previous exposure to written Eastern Dan were taught various combinations of tones and segments in parallel groups and their acquired skills were tested in dictation and oral reading tasks. The results point to an advantage for the experimental orthography that combines the punctuation tone marking strategy with biunique segmental correspondence and spe-

The field phase of this research project was supported by a public grant overseen by the French National Research Agency (ANR) as part of the “Investissements d’Avenir” program (reference: ANR-10-LABX-0083). This paper benefitted from discussions following its presentation at the /gɾafematik/ conference in Brest, France, 14–15 June 2018, a video of which can be viewed at <https://www.youtube.com/watch?v=W5Pir6cPoQs> (accessed 4 May 2019).

David Roberts

Independent linguistics, literacy and education consultant, working in collaboration with SIL International. B.P. 57, Kara, Togo

rbrdvd@gmail.com

Dana Basnight-Brown

Associate Professor and Research Scientist, Center for Cognitive and Developmental Research, United States International University - Africa. PO Box 60875, City Square, 00200, Nairobi, Kenya

dana.basnightbrown@gmail.com

Valentin Vydrin  0000-0002-7822-4173

Professor at INALCO and researcher at LLACAN-CNRS, Paris. LLACAN - UMR 8135 CNRS, 7 rue Guy Môquet - BP 8, 94801 Villejuif Cedex, France

vydrine@gmail.com

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1.

Fluxus Editions, Brest, 2019, p. 293–327. <https://doi.org/10.36824/2018-graf-robe>

ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

cial characters for marking vowels. Nevertheless, the language community has recently adopted Vydrin's reform in its entirety.

Abbreviations. C: consonant; H: high tone; L: low tone; M: mid tone; T: tone; TBU: tone bearing unit; V: vowel; xH: Extra-high tone; xL: Extra-low tone.

1. Background

1.1. Orthography Development

Dan¹ is a South Mande Niger-Congo language spoken in the Man, Danané and Biankouma prefectures of the Montagnes district of Côte d'Ivoire, and in Liberia where it is called Gio; there are also some Dan villages in Guinea. Since Dan has over forty dialects in Côte d'Ivoire alone, the decision was made in the 1970s to develop two varieties in this country: Western Dan (based on the Blo dialect) and Eastern Dan (based on the Gweetaa dialect, and including the dialect of Man, the main population centre of the entire Dan population). It is the latter variety that is the subject of the present research. Dan is spoken by about 1,600,000 people in all three countries (Vydrin, 2016a), of whom it is estimated that 650,000 are Eastern Dan speakers (Eberhard, Simons, and Fennig, 2019).

Eastern Dan is an overwhelmingly oral society. The only language of instruction at school is French, the official language, and L1 literacy is consigned to informal adult education. The average Dan speaker cannot read or write his or her own language. Literacy classes were extremely numerous in the 1970s-1980s (Bolli 1980, p. 7; 1983, p. 3; Thomas 1978, pp. 1, 14, 16) but numbers have declined steeply since then.

Eastern Dan is unusual among African languages in that it has five level tones and six contours. This, combined with the fact that it is highly monosyllabic and isolating in its root structures, results in a language with an exceptionally heavy functional load of tone. In the 1970s, SIL researchers were faced with the unenviable challenge of developing a tone orthography for Eastern Dan while working with the limitations of manual typewriters. The solution they came up with was an orthography that marks tone fully² using word initial and word final punctuation

1. ISO 639-3: dnj. Exoglossonym: Yacouba.

2. We use the term *full tone marking* to refer to orthographic representations containing one symbol fewer than the number of contrastive level tones in the language. We reserve the term *exhaustive tone marking* for orthographies that mark each and every tone, a tradition that is virtually unknown in Africa but is not uncommon in Asia and Latin America.

marks:³ a radical departure from the traditional strategy of using superscript accents (Bolli, 1978). This orthography, first developed in 1974, survived a government imposed segmental reform in 1982 and was still in use when we undertook this research project in 2017. Henceforth, it will be referred to as “The 1982 orthography”.

The punctuation strategy was hailed locally as a breakthrough at the time and was replicated in no fewer than fifteen Mande, Kru and Kwa languages in Côte d’Ivoire as well as being validated at national level by the *Institut de Linguistique Appliquée* (ILA, 1979). Although it has seldom been adopted beyond the borders of Côte d’Ivoire, it has nevertheless received some attention among writing systems researchers (Fricke-Kappers 1991; Kutsch Lojenga 1993, pp. 13–14; Kutsch Lojenga 2014, pp. 57–58; Roberts 2013, p. 91).

1.2. Previous Experimentation

In 2015, Roberts and Vydrin collaborated with a remote team of researchers working in five African countries to run a cross-linguistic classroom experiment the aim of which was to test the contribution of full tone marking to reading and writing fluency in the orthographies of ten Niger-Congo languages, including Eastern Dan.⁴

Across the ten languages, a total of 308 readers were recorded orally reading four previously unseen texts with and without tone marks. The results were measured for reading speed, accuracy and comprehension. The participants also added tone marks to the unmarked versions of the texts using pencil and paper, and we used these data to measure tone writing accuracy.

Analysis of the reading results indicates that, among the 57 Eastern Dan participants, the presence of punctuation marks to indicate tone does not contribute to gains in oral reading speed. Neither does it have a measurable impact on comprehension, contrary to some of the other languages. Granted, the punctuation marks do slightly reduce the number of errors in oral reading, but their average number is much higher in Eastern Dan than in any of the other languages, irrespective of whether tone is marked. As for the writing results, the average success rate in adding punctuation marks to unmarked texts was just over 60%, and only 3.5% of the participants scored over 90%. All this suggests not only that the 1982 tone orthography may not be doing its job effectively, but

3. In fact, as we will see, the 1982 orthography contains a mixture of punctuation and mathematical symbols, but in this paper, for the sake of brevity, we will refer to all of them as punctuation symbols.

4. The other languages were Elip, Mmala, Yangben (Bantu A62), Yoruba, Idaasha, Ife (Ede), Nateni, Mbelime and Tem (Gur).

also that other orthographic elements, such as the lack of segmental adherence to the phonemic principle may also be contributing to lack of fluency (Roberts, submitted).

The results are not altogether surprising given that we had already identified Eastern Dan as being an outlier on four accounts. First, it has a far heavier functional load of tone than any of the other nine languages. Second, the Eastern Dan orthography is the only one of the ten in which tone is represented by punctuation marks. Third, the literacy primer (Tiémoko, Déli Tiémoko, Bolli, and Flik, 1994) contains no dedicated tone lessons. Fourth, the literacy program was decimated by two civil wars in 2002–2007 and 2010–2011.

2. Comparing the 1982 and 2014 Orthographies

2.1. Introduction

All orthography stakeholders—literacy personnel, Bible translators, linguists, and writers among others—agree that Eastern Dan must mark tone fully because the functional load of tone is so exceptionally heavy. If the 1982 orthography was to be reformed, then, the question was not largely⁵ one of diacritic density;⁶ it was rather to do with the choice of symbols, and their position with relation to the orthographic word. An alternative orthography, developed by Valentin Vydrin in 2014, marks tone fully with superscript diacritics, eliminates consonant and vowel over-representation, and replaces unlauded vowels with special characters. Henceforth, it will be referred to as “The 2014 orthography”. It was introduced to Eastern Dan orthography stakeholders at two meetings in Man in September 2014 and January 2017.

The development of the 2014 orthography presented an ideal opportunity for a second experiment following on from the worrying results of the first that would specifically investigate the choice of symbol and position for tone marks in the 1982 orthography more closely. Such an experiment would also add novel perspective to the tone orthography literature, which tends to be dominated by experiments testing the parameters of diacritic density (Bernard, Mbeh, and Handwerker 2002;

5. We state ‘largely’ because the 2014 orthography marks tone on word medial feet whereas the 1982 orthography was incapable of this. As a result, it has a higher diacritic density, but only slightly so because most words have only one foot. See Section 2.4 for details.

6. In this study, we use the term “diacritic” to refer to both superscript accents and word-initial and word-final punctuation. Diacritic density is precisely measurable by calculating the number of diacritics as a percentage of the total number of orthographic TBUs in a natural text (Bird, 1999, p. 89).

Bird 1999) and orthographic depth (Mfonyam 1989; Roberts, Snider, and Walter 2016).⁷ In the following sections, we summarize the differences between the 1982 and 2014 orthographies.

2.2. Consonants

Table 2 compares the consonantal grapheme-phoneme correspondences in the 1982 orthography (Vydrin and Kességbeu, 2008) and the 2014 orthography (Vydrin, Zeh, and Gué, 2019).

TABLE 1. Consonantal grapheme-phoneme correspondences in the 1982 and 2014 Eastern Dan orthographies

	Phoneme	1982	2014
Voiceless stops	/p/		<p>
	/t/		<t>
	/k/		<k>
	/kp, kw/		<kp, kw>
Voiced stops	/b/		
	/d/		<d>
	/g/		<g>
	/gb, gw/		<gb, gw>
Voiceless fricatives	/f/		<f>
	/s/		<s>
Voiced fricatives	/v/		<v>
	/z/		<z>
Implosives	/ɓ/	<bh, m>	<bh>
	/ɗ/	<dh, n>	<dh>
Continuants	/l/	<l, r>	<l>
	/y/		<y>
	/w/		<w>

The 1982 orthography contains three cases of allophonic over-representation where spelling represents the surface form. First, the phoneme /ɓ/ is pronounced [m] preceding a nasal vowel and [ɓ] elsewhere; these sounds are spelled <m, bh> respectively. Second, the phoneme /ɗ/ is pronounced [n] preceding a nasal vowel and [ɗ] elsewhere; these sounds are spelled <n, dh> respectively. Third, the

7. To our knowledge, Duitsman (1986) is the only other researcher to have tested the Ivoirian punctuation system for marking tone. However, his intervention in Western Krahn lasted only 90 minutes, and some of the participants had prior knowledge of one of the alternatives being tested. Variables were not controlled for and no reference is made to statistical significance. All in all, the experiment design and reporting have such serious flaws that the results can tell us very little.

phoneme /l/ is pronounced [ɾ] following a coronal consonant and [l] elsewhere; these sounds are spelled <r, l> respectively, however, many writers spontaneously abandon <r> in favor of <l>. The 2014 orthography eliminates the graphemes <r, m>, and maintains the grapheme <n> only for the purpose of representing nasal vowels; in this way it maintains a biunique phoneme-grapheme correspondence.

2.3. Vowels

Table 2 compares the vocalic grapheme-phoneme correspondences in the 1982 orthography (Vydrin and Kességbeu, 2008) and the 2014 orthography (Vydrin, Zeh, and Gué, 2019).

TABLE 2. Vocalic grapheme-phoneme correspondences in the 1982 and 2014 orthographies

	Phoneme	1982	2014
Front unrounded oral	/i/		<i>
	/e/	<e, ɛ>	<e>
	/ɛ/		<ɛ>
	/æ/	<ɛa>	<æ>
Back unrounded oral	/u/	<ü>	<u>
	/ɤ/	<ö, ü>	<ɤ>
	/ʌ/	<ë>	<ʌ>
	/a/		<a>
Back rounded oral	/u/		<u>
	/o/	<o, v>	<o>
	/ɔ/		<ɔ>
	/ɒ/	<aɔ>	<œ>
Front unrounded nasal	/ĩ/		<in>
	/ẽ/		<ɛn>
	/æ̃/	<ɛan>	<æ̃n>
Back unrounded nasal	/ũ/	<ün>	<un/
	/ʌ̃/	<ɛ̃n>	<ʌ̃n>
	/ã/		<an>
Back rounded nasal	/ũ/	<un>	<un>
	/õ/	<ɔ̃n>	<ɔ̃n>
	/õ̃/	<aɔ̃n>	<œ̃n>
Velar nasal	/ŋ/	<ng>	<ŋ>

The velar nasal /ŋ/ is best analyzed as being a vowel with a restricted distribution (Vydrin and Kességbeu, 2008).⁸ The 1982 orthography writes it as <ng>, and the 2014 orthography as <ŋ>.

8. Alternatively, it can be interpreted as a syllabic nasal, occupying an intermediate position between a vowel and a consonant. It cannot be interpreted as a consonant

The 1982 orthography contains three cases of vowel over-representation for speakers of the Gweetaa dialect, though each of these pairs of allophones appear to be contrastive in other dialects, including that of Man. First, the phoneme /e/ is pronounced [ɪ] on a xH tone syllable and [e] elsewhere; these sounds are spelled <ɪ, e>, respectively. Second, the phoneme /ɤ/ is pronounced [ɥ]⁹ on a xH tone syllable and [ɤ] elsewhere; these sounds are spelled <ö, õ>, respectively. Third, the phoneme /o/ is pronounced [ɔ] on a xH tone syllable and [o] elsewhere; these sounds are spelled <v, o>, respectively. The 2014 orthography eliminates <ɪ, ö, v> from the alphabet in order to maintain a biunique grapheme-phoneme correspondence, although it is intended that these vowels could still be distinguished as /ɪ, ɥ, v/ in certain dialects where /ɪ, ɥ, ɔ/ have phonological status.

In addition, the 1982 orthography writes three other back unrounded vowels with umlauts, the graphemes <ü, ö, ë> representing the phonemes /ɯ, ɤ, ʌ/, respectively. Since superscript tone diacritics are not easily combinable with the umlauts, the 2014 orthography spells these three vowels with the characters <ɯ, ɤ, ʌ> respectively.

Two long open front vowels /ææ, ɒɒ/ also occur. However, it has only recently been discovered that their short counterparts /æ, ɒ/ also exist, albeit seldom (Vydrin, 2016b, p. 472). The 1982 orthography under-represents this length contrast, writing both the short and long vowels as <εa, aɔ>, respectively. The 2014 orthography represents the short vowels as <æ, œ>, and the long vowels as <ææ, œœ>, respectively.¹⁰

2.4. Tones

Eastern Dan has five phonemic level tones, extra high (xH), high (H), mid (M), low (L), extra-low (xL).¹¹ These can be combined in four falling contour tones and two rising contour tones. All falling tones finish at the xL level; both rising tones begin at the M level (Flik 1977; Vydrin and Kességbeu 2008, pp. 10–11).

The 1982 orthography uses punctuation symbols placed word initially and word finally to signal tone. Level tones are marked preceding

because /ɲV/ is unattested, no consonants bear tone, and no other nasal consonants are attested with which it might form a series.

9. Following Vydrin and Kességbeu (2008, p. 7), we use this symbol to indicate a near-close near-back unrounded vowel.

10. In addition, many words have free variation between /ææ ~ εε/ and /ɒɒ ~ ɔɔ/. The 1982 and 2014 orthographies both permit both spellings for these.

11. In Eastern Dan literacy classes, the two outermost tones are referred to as “very high” and “very low”.

the word. As for contour tones, the first element is marked word initially and the second word finally, but only on one-foot words (Kutsch Lojenga, 1993, ms 1989). The 2014 orthography, on the other hand, marks tones with superscript diacritics (Table 3).

TABLE 3. Grapheme-toneme correspondences in the 1982 and 2014 orthographies

Level tones	1982	2014	Contour tones	1982	2014
xH	/ǒ/	<"◊>	xH – xL	/ǒǒ/	<"◊◊>
H	/ó/	<'◊>	H – xL	/óǒ/	<'◊ǒ>
M	/ō/	<◊>	M – xL	/ōǒ/	<◊ǒ>
L	/ò/	<◊◊>	L – xL	/òǒ/	<◊◊ǒ>
xL ¹²	/ṽ/	<-◊>	M – H	/ōó/	<◊ó>
			M – xH	/ōǒ/	<◊ǒ>

Some consider the marking of L and xL tones as <◊◊, -◊> respectively to be counter-intuitive. For an explanation of the historical reasons for this choice, see Roberts (submitted).

By far the majority of Eastern Dan words have only one foot, and any words with three or more feet tend to be compounds. Since the 1982 orthography is incapable of marking tone on word medial feet there is a limited amount of under-representation on words of more than one foot.¹³

The 1982 orthography marks one symbol fewer than the number of phonemic tones in the language, representing M tone with absence of an accent. The 2014 orthography might have followed this principle, but it was considered more appropriate to represent all five phonemic levels, permitting the second vowel of a level sequence to be unaccented; thus, [ǒǒ] is spelled <ǒ◊>.¹⁴

12. The xL tone tends to be typed as <-◊> but handwritten as <_◊>.

13. In fact, the 1982 orthography has a greater degree of tonal under-representation than necessary, because contour tones on two-foot words, which could easily be represented with the punctuation system, are not fully marked for reasons that remain unclear. We did not specifically address this issue in our experiment.

14. In a limited number of words, a single short vowel bears a HxL contour which the 2014 orthography represents with a circumflex ([ó̂]; 1982 <'◊->; 2014 <ó̂>). As for the even less frequent MxL contour, all the affected words fortuitously contain a nasal vowel, so the 2014 orthography writes them without introducing an extra diacritic (e.g., ([dĩ̃]); 1982 <din->; 2014 <dĩ̃> *bunger*). Both these contours were excluded from the experimental teaching materials on account of their extreme infrequency.

2.5. Summary

Table 4 summarizes the consonant, vowel and tone representations in the 1982 orthography and the 2014 orthography.

TABLE 4. Summary of the 1982 and 2014 orthographies

	1982	2014
Consonants	over-representation of 3 consonants.	biunique correspondence; maintains <n> but only to mark nasal vowels.
Vowels	over-representation of 3 vowels; limited use of special characters; 2 oral vowels written as digraphs.	biunique correspondence; replaces umlauted vowels with special characters; replaces digraphs with special characters.
Tone	largely biunique correspondence, but some under-representation of words of more than one foot; punctuation in word initial and final position.	biunique correspondence; accents in superscript position.

3. Arguments for and against Orthography Reform

Before proceeding with an account of the experiment, it will be helpful to discuss various arguments for and against orthography reform that we have noted during the course of our fieldwork. These will be framed within Smalley's (1963) five criteria for developing optimal orthographies: maximum motivation and acceptance¹⁵ (Section 3.1), maximum representation of speech (Section 3.2), maximum ease of learning (Section 3.3), maximum ease of transfer (Section 3.4), and maximum ease of reproduction (Section 3.5).

3.1. Maximum Motivation and Acceptance

This criterion has to do with the extent to which learners are motivated to use the orthography, and its acceptance by society and those in au-

15. The original states: "Maximum motivation for the learner, and acceptance by his society and controlling groups such as the government." (Smalley, 1963, p. 34)

thority. Orthography development typically takes place over decades and centuries rather than months and years, so a reassessment by the second generation of literacy stakeholders should be viewed as a perfectly acceptable stage in a longer process (Karan, 2014). Yet some question why there is a need to change when teachers have successfully taught the 1982 orthography for so long. The older ones among them well remember that the literacy program took years to recover from the only previous experience of reform, in 1982, and that was at a time when motivation for literacy was fervently high. They fear that a second reform may have a much greater negative impact on a generation in which motivation for literacy is much lower. A counter-argument is that any reform may prove to be less turbulent than it was in 1982, precisely because it will impact far fewer people.

Furthermore, some consider that the scope of the 2014 orthography is too far-reaching, because it involves multiple changes to all three phonological levels: consonants, vowels and tone. Recent orthography reform in European languages warns us that resistance is likely towards even the most modest and conservative changes. A literate community develops an attachment to a familiar orthography, gradually becoming blind, or even attached to its imperfections. To outsiders, the cumulative visual effect of the punctuation symbols in the 1982 orthography can look cluttered and aesthetically displeasing. Yet Eastern Dan learners never complained of this; on the contrary, they were proud of its distinctiveness.

3.2. Maximum Representation of Speech

This criterion has to do with the extent to which the orthography adheres to the phonemic principle. The 1982 orthography over-represents some consonants and vowels, and under-represents tone on words of more than one foot. A hard-nosed linguist might go further, arguing that the back unrounded vowel series contains an illogical mixture of symbols: Four of them are written with umlauts <ü, ö, ø, ë>, but the fifth <a> is not. Also, the umlauted letters <ü, ö, ø> are graphic modifications of the back rounded vowels <u, v, o>, whereas the umlauted letter <ë> is a graphic modification of the front unrounded vowel <e>, and in any case the graphemes <e, ë> do not share the same aperture. However, such picky linguistic concerns are generally far removed from the needs of learners. The 2014 orthography resolves all these issues by adhering to the phonemic principle and being relatively consistent with the IPA.

3.3. Maximum Ease of Learning

This criterion has to do with the extent to which the orthography is easy for learners to master. Some have expressed concern that the allophonic

representation of the three vowel phonemes /e, o, ɤ/ on xH tone syllables with the graphemes <ɛ, ɔ, ɛ>, respectively takes up too much time in the classroom. Teachers of the 1982 orthography even present the letters <ɛ, ɔ> to pupils as “i malade” and “v malade” (“sick i” and “sick v”) respectively because of their wobbly shapes, which would seem to indicate that they are denigrating them. The 2014 orthography eliminates this allography, but not entirely satisfactory, because recent research has revealed that the series /ɪ, ʊ, ʏ/ are phonemic in at least some dialects. A further pedagogical issue is that the under-representation of tone on words with more than one foot in the 1982 orthography leads many learners to avoid compounding which would otherwise be helpful for word identification. Again, the 2014 orthography eliminates this problem.

3.4. Maximum Ease of Transfer

This criterion has to do with the extent to which the orthography facilitates transfer of literacy skills to and from other languages. Some stakeholders have expressed concern that the allophonic representation of the consonant phonemes /b, d, l/ with the graphemes <m, n, r> in the 1982 orthography places an unnecessary pedagogical burden on teachers and learners. But these letters were included out of a concern that those who are literate in French will be used to hearing and writing these sounds. The 2014 orthography eliminates this allography, thus tending more towards the needs of monolingual learners.

Another transfer issue has to do with regional practice. At least two of the fourteen Ivoirian languages that used to use the punctuation strategy for marking tone—Mwan (Perekhval'skaya and Yegbé, 2018) and Guro (N. Kuznetsova, O. Kuznetsova, and Vydrin, 2009)—have abandoned it, and plans are afoot to switch in Western Dan (Loh Japhet p.c.) and Gban too (Taki Oya Robert p.c.). Toura has also recently replaced the digraph <ng> with the special character <ŋ> (Thomas Bearth, p. c.). The ILA is in favor of these changes. However, although government authorities might give high priority to inter-language harmonization as a sign of national unity, in practice few Ivoirians learn to read and write in more than one local language.

3.5. Maximum Ease of Reproduction

This criterion has to do with the extent to which the orthography facilitates typing and publishing. One of the main advantages of the 1982 orthography was that the four punctuation symbols required were all available on manual typewriters, a blind eye being turned to the potential for confusion of the L tone symbol with the equal sign <=> in math

booklets.¹⁶ But beyond this, both orthographies contain elements that pose challenges for reproduction: The 1982 orthography has four special characters and one (pervasive) diacritic; the 2014 orthography has seven special characters and seven diacritics.

Neither could the developers of the 1982 tone marking strategy have foreseen that it would one day throw up numerous drawbacks in the early days of computer use. First, it did not facilitate alphabetical sorting, since words are merely grouped according to their word-initial tone marks. Second, spreadsheet programs interpret the L tone symbol <=> as introducing a mathematical formula, and the H tone symbol <'> as introducing a string of text. Third, word processing programs interpreted the punctuation marks as being beyond the domain of the orthographic word, so they were excluded from operations such as word selection and searches. The xL tone symbol <-> was particularly problematic because software interpreted it as a hyphen, triggering unwanted line-breaks: A randomly picked 23-line article from the *Pamebbame* newspaper contains no fewer than seven cases of this. Such issues are by no means insurmountable in the era of Unicode,¹⁷ as long as writers are trained to choose word-forming characters—i.e., modifier letters that resemble the standard but are endowed with word-forming properties (Cahill 2019, p. 4; SIL 2018, pp. 5–6¹⁸)—but most Eastern Dan literates remain unaware of this and resort to the simple keystrokes at their fingertips. As for the issue of alphabetic sorting, locale data—i.e., basic information on certain language specific needs and preferences that are necessary to display text including sort order (Osborn, 2010, p. 75)—could be submitted to Unicode, but very few African languages have done this to date.

In any case, the real IT challenge nowadays is ensuring that Eastern Dan is reproducible on smartphones, which are far more widespread among young people than computers ever were in their parents' generation. An Android keyboard for the 2014 orthography has already been developed for this purpose.¹⁹ For further discussion of Eastern Dan IT compatibility, see Paterson III (2019).

16. The use of double quotation marks in the 1982 orthography never created a conflict with symbolizing direct speech, because it follows the French convention of «chevrons» for this purpose.

17. www.unicode.org/versions/Unicode12.0.0/ (accessed 7 May 2019). See also Anderson, R., and Whistler (2005). For a summary of Unicode and the background to its development, see Osborn (2010).

18. Specifically: <"> U+02BA MODIFIER LETTER DOUBLE PRIME; <'> U+02B9 MODIFIER LETTER PRIME; <-> U+02D7 MODIFIER LETTER MINUS SIGN or <-> U+2011 NON-BREAKING HYPHEN (all introduced in Unicode 1.1, 1993); <=> U+A78A MODIFIER LETTER SHORT EQUALS SIGN (introduced in Unicode 5.1, 2008).

19. We acknowledge Andrew Cunningham's work in developing the freely downloadable Eastern Dan Android keyboard. Users should install Keyman (<https://keyman.com>) then follow this link: <https://drive.google.com/>

Budgetary considerations are not to be ignored in a social context with extremely limited financial resources. It would be irresponsible for an outsider to promote orthography reform without also finding ways of financing the reproduction of literature and the organization of transition classes. It is incontrovertible that texts written in the 1982 orthography, with its linear tone marking, are 10% longer and therefore considerably more expensive to publish than those written using the 2014 orthography. However, some have expressed concern that the costs of reprinting the existing literature would be prohibitive and that literature will henceforth be split into pre- and post-reform publications. Others are of the opinion that, if reform must happen, it should be before the publication of the whole Bible (planned for 2020) because once it is in print, it will become authoritative, on the evidence that the New Testament (SBI, 1991) has proved to be by far the best-selling Eastern Dan book.

The above arguments only put forth the possible positive and negative consequences of spelling reform. In the following sections, we complement this qualitative approach with quantitative data from the classroom investigating the effects of the 1982 and 2014 orthographies on reading and writing performance.

4. The Experiment

4.1. Aim

The aim of the experiment was to test two ways of marking:

- (i) *tone*: word initial and final punctuation (the 1982 orthography) against superscript diacritics (the 2014 orthography);
- (ii) *segments*: over-representation of consonants and vowels (the 1982 orthography) against biunique grapheme-phoneme correspondence (the 2014 orthography), and umlauts (the 1982 orthography) against special characters (the 2014 orthography).

The experiment tested oral reading (measured in terms of speed, accuracy and comprehension) and writing (measured in terms of accuracy).

4.2. Design

The experiment followed a between-groups 2×2 factorial design, permitting us to examine the effects of segments and tone independently

of each other, as well as any potential interaction between them. Participants were divided into four parallel classes by matched random assignment, and each experimental group was taught one of four orthographies in an intensive five-day course. Participants were given eleven dictation tests during the intervention and oral reading tasks following it. Table 5 shows the overall design and the number of participants in each experimental group. Orthography A is the 1982 orthography; orthography B combines the 2014 segments with the 1982 tones; orthography C combines the 1982 segments with the 2014 tones; orthography D is the 2014 orthography.

TABLE 5. Experiment design

		SEGMENTS	
		1982	2014
TONE	1982	A (<i>n</i> = 16)	C (<i>n</i> = 17)
	2014	B (<i>n</i> = 17)	D (<i>n</i> = 18)

Figures 1–4 illustrate the visual effect of the four orthographies.²⁰

"Yua -ya =Göö- 'kun, "kεε "yua 'ö =Göö- -bha, mε 'bha 'yaa -a dɔ. =Göö- zuë" -ya -kě, =Göö- "ting -yö -sü, -a suë" -nu =wa 'go mü. =Göö- 'yaa wlüü" -dhe yö, 'yaa ö 'bhuëë- bho, -a 'bhuëë- =ya -da. =Wa =Göö- zü zua, "yua 'yaa bo. =Wa "bhuëë kö bho =dua 'ka, =wa -a -kpa, =wa -a "yi nu =Göö- -dhe, -a -bha "yua 'yaa bo.

FIGURE 1. Orthography A: 1982 SEGMENTS, 1982 TONES

Yúa yà Göö kún, kée yúa ó Göö bhà, mē bhá yáa à dɔ. Göö züë yà kě, Göö tǐng yò sǔ, à sǔë nù wà gó mü. Göö yáa wlüü dhè yö, yáa ö bhúëë bhō, à bhúëë yà dǎ. Wà Göö zǔ zúa, yúa yáa bō. Wà bhúëë kō bhō dùa ká, wà à kpà, wà à yí nù Göö dhè, à bhà yúa yáa bō.

FIGURE 2. Orthography B: 1982 SEGMENTS, 2014 TONES

20. English translation of the text sample: "Geu has grown sick, but nobody knows what kind of illness it is. Geu has heartache, difficulty breathing, and his fingernails have fallen off. Geu cannot stand up, he does not shave himself, and his beard has grown. He has had an injection in the buttocks, but the illness has not stopped. People have cut cashew tree shavings with an axe, boiled them and given the concoction to Geu, but his illness has not stopped."

"Yua =ya =Gɔɔ- 'kun, "kɛɛ "yua 'ɔ =Gɔɔ- -bha, bhɛn 'bha 'yaa -a dɔ. =Gɔɔ- zuɔɔ" -ya -kɔ, =Gɔɔ- "teŋ -yɔ -su, -a suɔ" -dhun =wa 'go bhun. =Gɔɔ- 'yaa wluu" -dhe yɔ, 'yaa ɔ 'bhuɔɔ- bho, -a 'bhuɔɔ- =ya -da. =Wa =Gɔɔ- zu zua, "yua 'yaa bo. =Wa "bhuɔɔ kɔ bho =dua 'ka, =wa -a -kpa, =wa -a "yi dhun =Gɔɔ- -dhe, -a -bha "yua 'yaa bo.

FIGURE 3. Orthography C: 2014 SEGMENTS, 1982 TONES

Yúa yà Gɔɔ kún, kɛɛ yúa ɔ Gɔɔ bhà, bhɛn bhá yáa à dɔ. Gɔɔ zúú yà kɔ, Gɔɔ tɛŋ yɔ sù, à sùú dhùn wà gó bhùn. Gɔɔ yáa wluú dhè yɔ, yáa ɔ bhúúú bhò, à bhúúú yà dà. Wà Gɔɔ zú zúa, yúa yáa bhò. Wà bhúúú kɔ bhò dùa ká, wà à kpà, wà à yí dhùn Gɔɔ dhè, à bhà yúa yáa bò.

FIGURE 4. Orthography D: 2014 SEGMENTS, 2014 TONES

4.3. Materials

Before the field phase, we prepared four versions of the experimental pedagogical and test materials. The lessons were based on those in the primer that teaches the 1982 orthography (Tiémoko, Déli Tiémoko, Bolli, and Flik, 1994) with the addition of dedicated tone lessons. The total number of lessons was reduced from 53 to 38 by eliminating those focusing on sound-symbol correspondences that the participants would already recognize from their knowledge of French. The four courses were identical in structure and content except with regard to the orthographies themselves.²¹

4.4. Timetable and Personnel

The field phase of the experiment, which took place in Man²² spanned twelve weeks from January to March 2017 (Table 6).

21. It should be noted that, because of time constraints, the experiment focused only on short words even though one of the main advantages of the 2014 orthography is that it permits tone marking on long words (See Section 2.4). Testing these was beyond the scope of our research.

22. We had originally planned to run the experiment in Santa (about 65 kms north-west of Man), where the reference dialect of Gweetaa is spoken. However, an exploratory visit there in September 2015 caused us to abandon this idea, because neither primary school had enough pupils, the education level of potential teachers was insufficient, and communication networks were unreliable. We then arranged to run the experiment at the Mont Glas Primary School in Man, but a nationwide teacher's strike in January 2017 forced us to abandon this plan only a few days before we were due to begin.

TABLE 6. Timetable for the field phase

	Week	Event	AM	PM
	1	Arrival, administration		
1st cycle	2	Supervisor training	A	C
	3–4	Pilot test	A	C
	5	Intervention	A, C	A, C
	6	Recordings, scoring	A, C	A, C
2nd cycle	7	Supervisor training	B	D
	8–9	Pilot test	B	D
	10	Intervention	B, D	B, D
	11	Recordings, scoring	B, D	B, D
	12	Administration, departure		

4.4.1. Supervisors

The principle author trained two supervisors in orthographies A and C in week 2 and in orthographies B and D in week 7.

4.4.2. Pilot Test Participants

The supervisors then led two, ten-day pilot tests (weeks 3–4, orthographies A and C; weeks 8–9, orthographies B and D) with a total of 18 adult participants. The aim of this phase was to test the experimental materials with a small manageable group, but in fact the results proved sufficiently trustworthy that we decided to integrate them into the final statistical analysis, controlling for this difference. From each group, we recruited one person as a classroom assistant and scorer for the main intervention.²³

4.4.3. Intervention Participants

The intervention itself took place in weeks 5 (orthographies A and C) and 10 (orthographies B and D) with a total of fifty adults. This was followed by recorded tasks and scoring in weeks 6 and 11 respectively. All participants, in the pilot tests and intervention, were L1 speakers of Eastern Dan and participated for payment. None of them had prior knowledge of either the 1982 or the 2014 orthographies, but all of them had a minimum of four years formal education, which meant that they

23. Our original plan was for this phase of the experiment to be a teacher training course from which we would recruit people to independently teach during the intervention, but two weeks proved to be insufficient time to achieve this objective.

were all minimally literate in French.²⁴ The language of instruction was Eastern Dan.

4.5. Intervention

Each course consisted of 38×30 minute content lessons, five revision lessons and eleven dictation tests. The total teacher-pupil contact time was 32.5 hours. Table 7 summarizes the lesson content.

TABLE 7. Lesson content

Lesson	Content	Lesson	Content
1	Short and long vowels	12	Revision: Contour tones
2	Level tones xH ~ H	13	/kp ~ gb/
3	Level tones H ~ M	14	/l/ [l, r]
4	Level tones M ~ L	15	/e/ [e, ɛ]
5	Level tones L ~ xL	16	/o/ [o, u]
6	Revision: Level tones	17	/ʁ/ [ø, ʊ]
7	/ŋ/	18–23	Oral vowels: /ʌ, u, ε, ɔ, æ, ɒ/
8	Falling contours xH-xL, H-xL	24–30	Nasal vowels: /ã, â, û, ê, ɔ̃, î, ü/
9	Falling contours H-xL ~ M-xL	31	/bh/ [bh ~ m]
10	Falling contours M-xL, L-xL	32	/dh/ [dh ~ n]
11	Rising contours M-xH, M-H	33–38	Diphthongs: /ia, iʌ, iʁ, ua, uʌ, uʁ/

4.6. Independent Variables

All the participants filled in a sociolinguistic questionnaire in French before the course began. Any whose L2 literacy skills were not sufficiently developed to do this were interviewed in Eastern Dan and responses recorded in French on their behalf. We also tracked lateness and absences. Table 8 summarizes the demographic variables.

24. The average age of participants in this experiment (28 years old) was much lower than in the first experiment (47 years old), because we proactively recruited young people out of a concern that little is being done to pass Eastern Dan literacy on to the younger generation. This was not an obligatory feature of the experiment design.

TABLE 8. Demographic variables

STATUS	Whether the participant attended the pilot test or the main intervention
GENDER	Participant's gender (male or female)
AGE	Participant's age (measured in years)
EDUCATION	Formal education completed (measured in years)
DIALECT	Participant's dialect profile
DIASPORA	How long the participant had spent living outside of the Eastern Dan territory, measured in years.
ABSENCE	Lateness and absences, measured in minutes.

This demographic data enabled us to assign participants in matched quadruplets. One-way ANOVAs conducted on all demographic variables retrospectively showed that the groups were indeed matched (e.g., AGE $F(4, 64) = .110$, $p = .95$; EDUCATION $F(4, 64) = .203$, $p = .90$).

4.7. Performance Variables

Tables 9 and 10 summarize the Eastern Dan and French performance variables associated with the dictation and oral reading tasks. Following Roberts (2013, 4, fn. 5), we use the term *Orthographic tone bearing unit* (TBU) to mean “any letter which can potentially be marked with a tone diacritic”.

TABLE 9. Eastern Dan performance variables

L1 DICTATION	Correct as a percentage of total number of words
L1 LIST SPEED	Oral reading speed of Eastern Dan word list measured in orthographic TBUs per minute
L1 LIST ERRORS	Errors per 100 orthographic TBUs on oral reading of Eastern Dan word list
L1 TEXT SPEED	Oral reading speed of Eastern Dan text measured in orthographic TBUs per minute
L1 TEXT ERRORS	Errors per 100 orthographic TBUs on oral reading of Eastern Dan text
L1 COMPREHENSION	Correct answers out of ten to comprehension questions about the Eastern Dan text.

TABLE 10. French performance variables

L2 TEXT SPEED	Oral reading speed of French text measured in syllables per minute
L2 TEXT ERRORS	Oral reading errors per 100 syllables on French text
L2 COMPREHENSION	Correct answers out of ten to comprehension questions about the French text.

5. Results

We ran a Multivariate Analysis of Covariance (MANCOVA) model with all dependent variables, examining them separately in terms of the main effect of segments and tones, and also the interaction between the two. Covariates entered into the model consisted of GENDER, AGE, EDUCATION, and all French performance variables (i.e., L2 TEXT SPEED, L2 TEXT ERRORS, L2 COMPREHENSION). The statistical analysis was performed using IBM SPSS software.²⁵ Only results that are significant ($p < .05$) and marginally significant ($p < .10$) are reported in the following sections.

5.1. Dictation

Eleven dictations consisting of 15 monosyllabic words each were spread across the five days of the intervention (165 words in total). Each dictation tested skills acquired in the immediately preceding lessons. The teacher said each word three times, preceding the first utterance of each triplet with consecutive cardinal numerals to provide a tone frame for the test word itself. The teacher repeated the entire list at the end. Dictation performance was first measured in terms of overall success, the whole word being scored as either correct or incorrect. Raw scores were converted to percentages (L1 DICTATION).²⁶

Table 11 reports mean accuracy rates on L1 DICTATION and standard deviations in parentheses. Orthography C (2014 SEGMENTS, 1982 TONES) emerges as the winner, while orthography B (1982 SEGMENTS,

25. <https://www.ibm.com/analytics/data-science/predictive-analytics/spss-trials> (accessed 21 October 2019).

26. We also scored results separately for consonants, vowels and tone, but subsequent ANOVAs revealed such high correlations between each pair (C and V: $r(68) = .91$, $p < .001$; C and T: $r(68) = .77$, $p < .001$; T and V: $r(68) = .91$, $p < .001$) that we decided to treat them as a composite in the MANOVA. The same comment applies to the oral reading error scores.

2014 TONES) lags far behind. This combination, with its diacritic stacking, is by far the most difficult to master.

TABLE 11. Overall dictation success rates (and standard deviations)

		SEGMENTS	
		1982	2014
TONE	1982	A: 63.71% (18.83)	C: 64.67% (19.56)
	2014	B: 43.17% (23.62)	D: 62.12% (17.16)

A preliminary ANOVA on L1 DICTATION revealed that those writing 1982 TONES (i.e., orthographies A and C) scored significantly higher than those writing 2014 TONES (i.e., orthographies B and D; $F(3, 64) = 5.71$, $p = .02$) and that those writing 2014 SEGMENTS (i.e., orthographies C and D) scored significantly higher than those writing 1982 SEGMENTS (i.e., orthographies A and B; $F(3, 64) = 4.24$, $p = .04$). There was also a marginally significant interaction between segments and tones ($F(3, 64) = 3.46$, $p = .07$) revealing that those writing 2014 SEGMENTS (i.e., orthographies C and D) perform equally well irrespective of tone condition but those writing 1982 SEGMENTS (i.e., orthographies A and B) perform better when also marking 1982 TONES.

The MANCOVA analysis confirmed a significant main effect of tone on L1 DICTATION, with those writing 1982 TONES (i.e., orthographies A and C) scoring higher than those writing 2014 TONES (i.e., orthographies B and D; $F(1, 44) = 9.900$, $p < .01$). Only one independent variable, L2 COMPREHENSION, predicted L1 DICTATION scores ($b = .408$, $t = 3.48$, $p < .001$).

We also examined dictation success rates on the seven individual segments that were manipulated in the experimental orthographies. Table 12 shows the individual scores as a percentage of the number of occurrences, with the highest scores underlined.²⁷ In all cases, those writing the 2014 SEGMENTS (orthographies C and D) score higher than those writing the 1982 SEGMENTS (orthographies A and B).

We scored dictation performance on individual level tones in the same way (Table 13). In three cases (H, M, L) those learning the 2014 TONES combined with the 2014 SEGMENTS score highest (orthography D). For the other two tones (xH, xL), those learning the 1982 TONES and the 2014 SEGMENTS scored highest (orthography C), although not dramatically more so than those learning the 2014 TONES and the 2014

27. For the vowel phonemes, the scores combine oral, nasal, short and long vowels.

TABLE 12. Dictation success rates on individual segmental phonemes

ORTHO- GRAPHY	SEGMENTS	TONES	/b/	/d/	/l/	/ŋ/	/e/	/ɾ/	/o/
A	1982	1982	76.14	86.46	89.35	83.41	69.64	74.11	58.75
B	1982	2014	73.26	70.10	89.98	77.89	63.87	69.41	45.88
C	2014	1982	95.19	91.67	97.39	90.06	77.31	83.58	84.71
D	2014	2014	94.44	<u>98.15</u>	94.65	89.46	74.60	<u>79.85</u>	<u>92.22</u>

TABLE 13. Dictation success rates on individual level tones

ORTHO- GRAPHY	SEGMENTS	TONES	/ó/ xH	/ó/ H	/õ/ M	/ò/ L	/ò/ xL
A	1982	1982	73.68	67.71	65.10	54.69	68.42
B	1982	2014	55.73	41.18	51.72	31.99	52.79
C	2014	1982	85.14	60.78	72.55	58.82	74.15
D	2014	2014	<u>81.73</u>	<u>80.56</u>	<u>84.26</u>	<u>65.63</u>	<u>74.12</u>

TABLE 14. Dictation success rates on individual contour tones

ORTHO- GRAPHY	SEGMENTS	TONES	Falling			Rising		
			/óò/ xHxL	/óò/ HxL	/õò/ MxL	/õó/ LxL	/õó/ MxH	/õó/ MH
A	1982	1982	66.67	66.18	45.63	75.00	69.64	65.63
B	1982	2014	58.82	41.18	37.65	35.29	49.58	<u>29.41</u>
C	2014	1982	66.67	69.20	53.53	82.35	79.83	36.76
D	2014	2014	<u>75.93</u>	<u>68.95</u>	<u>63.33</u>	<u>55.56</u>	<u>65.87</u>	-

SEGMENTS (orthography D). Scores for orthography B (1982 SEGMENTS, 2014 TONE), again, are consistently the lowest.

A similar pattern emerges when scoring contour tones (Table 14). For three of the tones (H-xL, L-xL, M-xH), participants perform more accurately when writing the 2014 SEGMENTS combined with the 1982 TONES (orthography C). But for two of the tones (xH-xL, M-xL), it is orthography D (2014 SEGMENTS, 2014 TONES) that attracts the highest average scores.²⁸

28. The score for writing the MH contour tone in Orthography D is unavailable due to a data entry error. There were only four occurrences.

5.2. Oral Reading (List)

The intervention was followed by recorded oral reading tests spread over two consecutive days. First, each participant was recorded reading a list of 20 Eastern Dan one-foot words beginning with the phonemes /b, d/.²⁹ Speed was measured in terms of orthographic TBUs per minute (L1 LIST SPEED).

Table 15 reports the results of L1 LIST SPEED. Those reading the 1982 SEGMENTS and 1982 TONES (orthography A) read fastest, while those reading the 2014 SEGMENTS and 2014 TONES (orthography D) read slowest.

TABLE 15. Mean oral reading speed of Eastern Dan word list in orthographic TBUs per minute (and standard deviations)

		SEGMENTS	
		1982	2014
TONE	1982	A: 14.45 (6.01)	C: 10.29 (4.25)
	2014	B: 10.53 (6.12)	D: 7.38 (2.90)

The MANCOVA analysis revealed a significant main effect of segments on L1 LIST SPEED ($F(1, 44) = 13.274$, $p < .001$, partial $\eta^2 = .232$), with those reading the 1982 SEGMENTS ($M = 13.80$) performing faster than those reading the 2014 SEGMENTS ($M = 9.35$). It also revealed a marginally significant main effect of tones on L1 LIST SPEED ($F(1, 44) = 2.883$, $p < .10$, partial $\eta^2 = .061$), with those reading the 1982 TONES performing faster than those reading 2014 TONES.

As for errors, they were defined as substitutions, insertions and omissions, and did not include repetitions and self-corrections. Raw error counts were converted to errors per 100 orthographic TBUs (L1 LIST ERRORS). Table 16 reports the results of overall reading errors on the Eastern Dan word list. Those reading the 2014 SEGMENTS and 1982 TONES (orthography C) made the fewest errors, whilst those reading the 2014 SEGMENTS and 2014 TONES (orthography D) made the most.

The MANCOVA analysis revealed a marginally significant interaction for L1 LIST ERRORS ($F(1, 44) = 3.462$, $p < .10$, partial $\eta^2 = .0739$). For those reading 1982 SEGMENTS, it made little difference whether they were reading 1982 TONES (Orthography A) or 2014 TONES (Orthography B). However, for those reading 2014 SEGMENTS, also reading 2014

29. We focused on these two phonemes at the specific request of Valentin Vydrin who considers them to be essential to his reform.

TABLE 16. Mean L1 oral reading errors per 100 orthographic TBUs of Eastern Dan word list (and standard deviations)

		SEGMENTS	
		1982	2014
TONE	1982	A: 42.89 (19.46)	C: 39.78 (14.13)
	2014	B: 40.11 (17.88)	D: 48.24 (21.06)

TONES (Orthography D) resulted in significantly more errors than those reading 1982 TONES (Orthography C).

5.3. Oral Reading (Text)

In the same recording session, each participant was recorded orally reading two previously unseen texts, one in Eastern Dan (193 words), the other in French (143 words). Speed was measured in terms of orthographic TBUs per minute for Eastern Dan (L1 TEXT SPEED) and syllables per minute for French.³⁰

Table 17 reports the mean results of oral reading speed of the Eastern Dan text. Those reading the 2014 orthography (orthography D) read slower than the other three groups who all performed at a similar rate.

TABLE 17. Mean oral reading speed of Eastern Dan text in orthographic TBUs per minute (and standard deviations)

		SEGMENTS	
		1982	2014
TONE	1982	A: 32.82 (8.94)	C: 33.75 (9.11)
	2014	B: 33.66 (12.67)	D: 25.07 (6.67)

The MANCOVA analysis revealed a statistically significant interaction between segments and tones for L1 SPEED ($F(1, 44) = 4.341, p < .05$, partial $\eta^2 = .090$). For those reading 1982 SEGMENTS, whether they read 1982 TONES (orthography A) or 2014 TONES (orthography B) made little

30. We consider the classic ‘words per minute’ measure to be inappropriate for cross-linguistic comparison, because words vary in language between languages. For further discussion of this issue, see Roberts (submitted).

difference to their oral reading speed. However, combining 2014 SEGMENTS and 2014 TONES (orthography D) resulted in a much slower reading speed than combining 2014 SEGMENTS with 1982 TONES (orthography C).

Errors were defined as before. Raw error counts were converted to errors per 100 orthographic TBUs for Eastern Dan (L1 TEXT ERRORS) and errors per 100 syllables for French (L2 TEXT ERRORS). None of the differences between the four orthographic conditions was statistically significant.

5.4. Comprehension

The recording sessions also included two comprehension tasks, orally answering ten questions each about the Eastern Dan and French texts. In both cases, the questions were asked and answered in Eastern Dan. Questions were devised to test a mixture of explicit and implicit information (cf. Piper, Schroeder, and Trudell 2016, pp. 140–142), but scoring did not differentiate between these. Oral reading comprehension was measured in terms of correct answers out of ten (L1 COMPREHENSION, L2 COMPREHENSION). We found no statistically significant evidence that group assignment had an impact on oral reading comprehension of the Eastern Dan text. Participants understood the text equally well regardless of the orthography they were exposed to.

5.5. Summary

Table 18 summarizes the results of the statistical analysis, and shows that quantitative evidence falls uniquely in favor of orthography C on three of the eight measures, and partially so on two others. The experimental orthography that employs the 2014 segments but maintains the 1982 tone marking strategy is therefore the most efficient in promoting reading and writing fluency.

6. Discussion

6.1. Methodology

It will be helpful to comment on various aspects of the experiment design before interpreting the results.

The choice of sample was a compromise. On the one hand, it would arguably have been preferable to conduct the experiment with illiterates to avoid the possibility of any influence from French. On the other hand,

TABLE 18. Summary of the statistical analysis

Task	Evidence in favor of orthography...	Segments	Tone
Dictation accuracy (overall)	C	2014	1982
Dictation accuracy (individual segments)	C, D	2014	1982, 2014
Dictation accuracy (individual tones)	C, D	2014	1982, 2014
Oral reading speed (list)	A	1982	1982
Oral reading accuracy (list)	C	2014	1982
Oral reading speed (text)	C	2014	1982
Oral reading accuracy (text)	-	-	-
Oral reading comprehension (text)	-	-	-

working with adults with a minimum of formal schooling meant we did not have to teach the Eastern Dan alphabet from scratch in the limited time available; 12 of the 36 letters were already known.

The technique of preceding each word in the dictation task with consecutive cardinal numbers as tone frames proved effective, as participants would have been unable to identify the tones of words in isolation. A more authentic way of achieving the same outcome would be to embed the target word in a natural frame (e.g., “I saw a *noun*”; “I like to *verb*”), while still having participants write only the test word.

We have a lingering concern about lesson order. The fact that the implosive phonemes /b, d/ have nasal allophones [m, n] which are rendered explicit in two of the experimental orthographies left us with no choice but to teach them after the nasal vowels. Yet their high frequency in natural contexts would have been a reasonable argument for teaching them much earlier, and doing so would have had the benefit of greatly amplifying the stock of available words for the initial lessons. Furthermore, teaching the two implosives early on would have better prepared participants for the oral reading task which specifically focused on a list of words beginning with them. We did not control for lesson order, but it would be desirable to develop ways of doing so in future experiments.

With these methodological concerns in mind, we now turn to a discussion of the experiment results as they impacted writers and readers.

6.2. Writing Results

The results of this experiment show that, for writers of Eastern Dan, the punctuation strategy is easier to master than superscript diacritics

for marking tone. This is likely to be, at least in part, because of their linear position. It is often remarked that, in languages that mark tone with superscript diacritics, writers often formulate entire sentences before returning to fill in the diacritics, while others leave them out completely. These two writing practices have always been completely absent in Eastern Dan, because the position of the punctuation symbols forces the writer to make choices about tone marking simultaneously with those concerning consonant and vowel symbols.

As for consonants and vowels, Eastern Dan writers find the 2014 segments easier to write than the 1982 segments and the obvious explanation is that there are fewer symbols to master in the 2014 orthography. The experimental courses contained six lessons in which those teaching orthographies A and B had to introduce two symbols, while those teaching orthographies C and D could use the equivalent time to focus on one symbol. When over-representation is avoided it frees up teaching time. Another possible explanation for the advantage of the 2014 segments is that the 1982 segments contain four vowel graphemes written with umlauts, dramatically increasing the diacritic density which is already relatively high because of tone marking. The diacritic density of orthographies C and D, in which only tone is marked, is 57.3%, whereas that of orthographies A and B, in which tone and some vowels are marked is 92.4%. Writers make gains when the orthography steers clear of any potential for visual crowding.

As for orthography B, no Eastern Dan literacy stakeholder is suggesting it as a viable system. It is awkward typographically, because it superimposes tone diacritics on umlauted vowels. But including this permutation was necessary in order to achieve a balanced design, and it incidentally provided an opportunity to test the effect of stacked diacritics. The low scores for Orthography B suggest that they should be avoided in orthography design.

6.3. Reading Results

An orthographic strategy that benefits writers does not necessarily produce equivalent advantages for readers. In the reading tasks, the only statistically significant main effects are for reading speed of the list and the text, not for errors or comprehension. Neither the 2014 segments nor the 2014 tones helped participants to read the word list faster: they performed best with the 1982 orthography. However, once words are placed in context, a different pattern emerges: for those reading the text with the 2014 segments, combining these with the 1982 tones was more advantageous in terms of reading speed than combining them with the 2014 tones.

As for oral reading error rates, the 2014 tones increase the error rates when coupled with the 2014 segments on the word list. However, this effect was not replicated when reading the text. Crucially, the support of context enables readers of all four experimental orthographies to read with equal levels of accuracy. The same is true of comprehension: no particular orthographic variation perturbs the reader's understanding once words are placed in context. This is one of the most unexpected findings of the experiment and it stands as a reminder of the extent to which readers, at least those with pre-existing L2 literacy skills, can apparently adapt with ease to remarkably divergent orthographic strategies when transitioning to their L1, even one like orthography B that obviously has a less than optimal configuration.

7. Conclusion

The 1982 Eastern Dan orthography is well-known for its use of word-initial and word-final punctuation to mark tone, and discussions in the literature about this aspect of the orthography have tended to overshadow important segmental issues. Our results reveal that participants are struggling more with writing the 1982 consonants and vowels than they are with writing the 1982 tone marks.

The results of the writing, reading speed (text), and reading error (list) measures all point to an advantage for orthography C. The 1982 tone marks appear to be doing their job well, while the overrepresentation of consonants and vowels is clearly detrimental for writers and slows down reading speed. The 2014 segments have the social advantage that they could be introduced one by one over time, and there are also pedagogical implications. A revised literacy primer along the lines of orthographies C or D would contain six fewer segmental lessons, which would leave more room for incorporating designated tone lessons that are lacking in the existing primer.

Our experiment did not attempt to tease apart the parameters of symbolization (punctuation vs. diacritics) and position (word-initial and word-final vs. superscript). Therefore, if orthographies A and C are more effective than orthographies B and D, we still do not know whether it is because of the choice of symbols or because of the choice of position. This would make an interesting subject for future experimentation.

We found no convincing evidence that readers and writers are struggling with the counter-intuitive symbolization for L and xL tones. However, inverting them in the 1982 orthography would be desirable for two reasons. Pedagogically, it would enhance their iconic value, making them easier to teach; sociolinguistically, it would bring Eastern Dan into alignment with the 1979 government guidelines and practice elsewhere in the country.

However, such punctilious concerns have been unexpectedly swept away, as our research project ends with a curious twist. Even though the experimental results point in favor of orthography C, a recent meeting of 68 orthography stakeholders in Man on 8 December 2018 decided unanimously to adopt Vydrin's spelling reform (i.e., orthography D) in its entirety (Zeh, 2018, p. 2). Since then, five, two-week transition classes have been organized, retraining about 250 literacy workers (Emmanuel Zeh, p.c.). Several books have been published in the 2014 orthography, including a guide (Vydrin, Zeh, and Gué, 2019), a transition guide (Anonymous, 2019) and reading materials (Saint-Exupéry 2019; Tiémoko 2019).

Decision makers were doubtless influenced by the fact that the *Institut de Linguistique Appliquée* is now advocating the representation of tone by means of superscript diacritics in place of punctuation for Ivoirian languages. The linguistic arguments, such as the ability to mark tone on word medial feet, as well as those to do with IT compatibility, also contributed to consensus building. But the local enthusiasm for reform also suggests that the social process of being involved in a classroom experiment, with its opportunity for exposure to the 2014 orthography, has had a greater impact on decision makers than the scientific results of it. Any researcher involved in the process of such reforms should not underestimate the challenges of conveying complex quantitative experimental results to lay people who are empowered to reform their own orthography but whose cultural and educational background mean that they are not necessarily going to be persuaded by the scientific method.

Acknowledgements

We are grateful to David Share for his insightful advice about experiment design, to Teresa Heath for allowing us access to the SIL Abidjan archives, to Thomas Bearth and †Margrit Bolli for illuminating discussions about Eastern Dan orthography development and the ethno-literacy context, to Emmanuel Zeh and Josephine Kpan for their tireless help in organizing and running the experiment, to Mike Cahill, Dave Rowe and Hugh Paterson III for their helpful comments on a draft of this paper, and to the Focolari community in Man for their generous hospitality.

References

- Anderson, D., McGowen R., and K. Whistler (2005). "Unicode Technical Note #19: Recommendations for Creating New Orthographies". <http://www.unicode.org/notes/tn19/tn19-1.html>.

- Anonymous (2019). *Cours éclair pour les lecteurs du français apprenant à lire le dan de l'Est dans la nouvelle orthographe*. Man: Pàbhēnbhàbhèn.
- Bernard, Russell H., George N. Mbeh, and W. Penn Handwerker (2002). "Does Marking Tone Make Tone Languages Easier to Read". In: *Human Organisation* 61, pp. 339–349.
- Bird, Steven (1999). "When Marking Tone Reduces Fluency: An Orthography Experiment in Cameroon". In: *Language and Speech* 42, pp. 83–115.
- Bolli, Margrit (1978). "Writing Tone with Punctuation Marks". In: *Notes on Literacy* 23, pp. 16–18.
- (1980). "Yacouba Literacy Report 2 (March 1977–February 1979)". In: *Notes on Literacy* 31, pp. 1–14.
- (1983). "The Victor Hugoes in Dan Country: Developing a Mother-Tongue Body of Literature in a Neoliterate Society". In: *Notes on Scripture in Use* 5, pp. 3–14.
- Cahill, Michael (2019). "Marking Grammatical Tone in Orthographies: Issues and Challenges". In: *50th Annual Conference on African Linguistics*. University of British Columbia.
- Duitsman, John (1986). "Testing Two Systems for Marking Tone in Western Krahn". In: *Notes on Literacy* 49, pp. 2–10.
- Eberhard, David M., Gary F. Simons, and Charles D. Fennig, eds. (2019). *Ethnologue: Languages of the World*. 22nd ed. Online version: <http://www.ethnologue.com> (accessed 3 May 2019). Dallas, TX: SIL International.
- Flik, Eva (1977). "Tone Glides and Registers in Five Dan Dialects". In: *Linguistics* 201, pp. 5–59.
- Frieke-Kappers, Claertje (1991). *Working papers in linguistics*. Vol. 40: *Tone Orthography in African Languages—a Recommendation*. Amsterdam: Vrye Universiteit.
- ILA (1979). *Une orthographe pratique des langues ivoiriennes*. Abidjan: Institut de linguistique appliquée, Université d'Abidjan.
- Karan, Elke (2014). "Standardization: What's the Hurry". In: *Developing Orthographies for Unwritten Languages*. Ed. by M. Cahill and K. Rice. Dallas, TX: SIL International, pp. 107–138.
- Kutsch Lojenga, Constance (1993). "The Writing and Reading of Tone in Bantu Languages". In: *Notes on Literacy* 19, pp. 1–19.
- (2014). "Orthography and Tone: A Tone System Typology with Implications for Orthography Development". In: *Developing Orthographies for Unwritten Languages*. Ed. by M. Cahill and K. Rice. Dallas, TX: SIL International, pp. 49–72.
- Kuznetsova, Natalia, Olga Kuznetsova, and Valentin Vydrin (2009). "Modeling a New Guro Orthography". In: *Proceedings of the 2nd Conference on Language Documentation and Linguistic Theory*. Ed. by P.K. Austin et al. London: SOAS, University of London, pp. 193–203.
- Mfonyam, Joseph Ngwa (1989). "Tone in Orthography: The Case of Ba-fut and Related Languages". Thèse d'état. Université de Yaoundé.

- Osborn, Don (2010). *African Languages in a Digital Age: Challenges and Opportunities for Indigenous Language Computing*. Cape Town: Human Sciences Research Council.
- Paterson III, Hugh (2019). "A Text Input Analysis of Eastern Dan". Masters dissertation. University of North Dakota.
- Perekhval'skaya, Elena and Moïse Yegbé (2018). "Dictionnaire mwan-français". In: *Mandenkan* 60, pp. 3–122.
- Piper, Benjamin, Leila Schroeder, and Barbara Trudell (2016). "Oral Reading Fluency and Comprehension in Kenya: Reading Acquisition in a Multilingual Environment". In: *Journal of Research in Reading* 39, pp. 133–152.
- Roberts, David, ed. (submitted). *Tone Orthography and Literacy: The Voice of Evidence in Ten Niger-Congo Languages*. Amsterdam: John Benjamins.
- (2013). "A Tone Orthography Typology". In: *Typology of Writing Systems*. Ed. by S.R. Borgwaldt and T. Joyce. Amsterdam: John Benjamins, pp. 85–111.
- Roberts, David, Keith Snider, and Stephen L. Walter (2016). "Neither Deep nor Shallow: Testing the Optimal Orthographic Depth for the Representation of Tone in Kabiye (Togo)". In: *Language and Speech* 59, pp. 113–138.
- Saint-Exupéry, Antoine de (2019). *Gblùdājgb̄-db̄án* [*The Little Prince, in Eastern Dan*]. Trans. by Nestor Gué, Valentin Vydrin, and Emmanuel Zeh. Man: Pābhēnhābhēn.
- SBI (1991). *Naw-sē 'ōgo Atanna "piü-a 'sēēdhe* [*The New Testament in Dan "Gweetaaww*]. Abidjan: Société Biblique Internationale et Association ivoirienne pour la traduction de la Bible.
- SIL (2018). "Best Practices Using Non-alphabetic Characters in Orthographies: Helping Languages Succeed in the Modern World". https://www.sil.org/sites/default/files/tone_and_unicode_issues.pdf.
- Smalley, William A. (1963). "How Shall I Write This Language?" In: *Orthography Studies: Articles on New Writing Systems*. Ed. by W.A. Smalley. London: United Bible Societies, pp. 31–52.
- Thomas, Paule (1978). *Alphabétisation en yacouba*. Abidjan: Institut de Linguistique Appliquée.
- Tiémoko, Baba Sébastien (2019). *Wón db̄. Kwēzlāan sÁadbēbē dānwò guí* [*There are things. Fairy tale book in Eastern Dan*]. Ed. by Nestor Gué, Valentin Vydrin, and Emmanuel Zeh. Man: Pābhēnhābhēn.
- Tiémoko, Baba Sébastien et al. (1994). =*Danɔw 'sēēdhe -wv pō -kɔ 'gweetaaww* [*Dan "gweetaaww Primer*]. Abidjan: SIL.
- Vydrin, Valentin (2016a). "Tonal Inflection in Mande Languages: The Cases of Bamana and Dan-Gweetaa". In: *Tone and Inflection: New Facts and New Perspectives*. Ed. by E.L. Palancar and J.L. Léonard. Vol. 296. Berlin: De Gruyter — Mouton, pp. 83–105.

- (2016b). “Дан язык [Dan Language]”. In: *Языки мира: Языки манде [Languages of the World: Mande Languages]*. Ed. by Valentin Vydrin [Валентин Выдрин] et al. St. Petersburg: Нестор-История [Nestor-Historia], pp. 469–583.
- Vydrin, Valentin and Mongnan Alphonse Kességbou (2008). *Dictionnaire dan-français (dan de l’Est) avec une esquisse de grammaire du dan de l’Est et un index français-dan*. 1st ed. St. Petersburg: Musée d’anthropologie et d’ethnographie, Académie des sciences de la Russie.
- Vydrin, Valentin, Emmanuel Zeh, and Nestor Gué (2019). *Guide de nouvelle orthographe dan de l’Est pour les apprenants familiers avec l’ancienne orthographe*. Man: Pābhēnhābhēn - EDILIS.
- Zeh, Emmanuel (2018). *Rapport de l’assemblée générale d’adoption de la nouvelle orthographe en langue dan-est le samedi 8 décembre 2018 à Man-Libreville*. Unpublished manuscript.

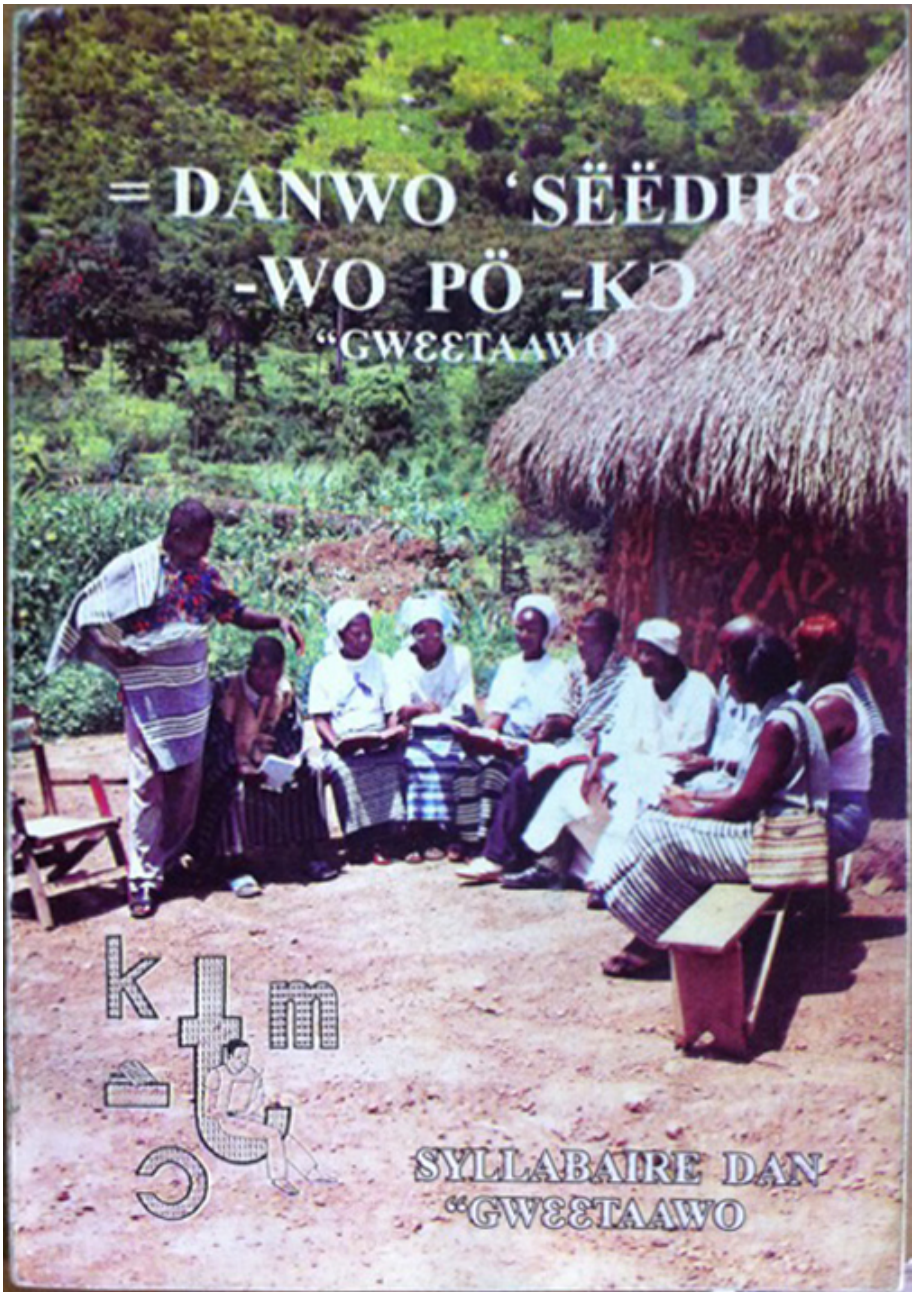


FIGURE 5. The cover of the Eastern Dan literacy primer in the 1982 orthography



FIGURE 6. An Eastern Dan teacher training class

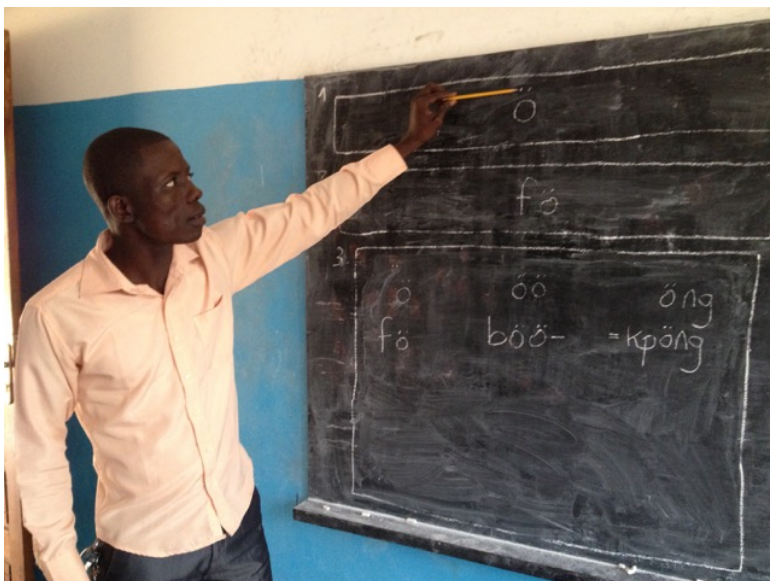


FIGURE 7. Emmanuel Zeh (literacy supervisor and principal collaborator) teaching the 1982 orthography



FIGURE 8. Pilot testing for the 2017 experiment

Antoine de Saint-Exupéry



Gblùdāgb̄-dhán

dànwò gú



Bháandhè

Pàbhēnbhàbhèn — EDILIS

2019

FIGURE 9. Cover of “The Little Prince” (Saint-Exupéry, 2019) translated into Eastern Dan, a new publication in the 2014 orthography

Malayalam Orthographic Reforms. Impact on Language and Popular Culture

Kavya Manohar & Santhosh Thottingal

Abstract. Malayalam is a language spoken in India, predominantly in the state of Kerala with about 38 million native speakers. The Malayalam script evolved from Brahmi through Grantha alphabet and Vattezhuthu writing systems. The script orthography has acquired its uniqueness with its complex shaped ligatures formed by the combination of consonants and vowel sign forms. The number of unique graphemes in this system exceeds 1,200. The orthographic styles were constantly evolving. In 1971 there was a Governmental intervention in the orthography, to reduce its complexity and to address the difficulties in typesetting and printing. This paper is an attempt to explore the impact of this orthographic reforms on various aspects of script usage including popular culture, media, textbooks, graffiti and handwriting. We will also analyse the impact of Unicode and the advancement in digital typography on the orthographic diversity of Malayalam script.

1. Introduction

With 38 million native speakers Malayalam is the official language of Kerala, in southern India. Malayalam used to be written in Vattezhuthu, the earliest known sample dating back to 830 AD. The modern Malayalam script evolved from Grantha alphabet which was a script for Sanskrit. Both Vattezhuthu and Grantha have their roots in the Brahmi script. As of today, Unicode has encoded 18 vowels and 37 consonants, some of them being archaic. Figure 1 illustrates the Malayalam code block as per Unicode 12.1¹.

Kavya Manohar  0000-0003-2402-5272

Assistant Professor

Department of Electronics and Communication Engineering

Government Engineering College, Palakkad, India

sakhi.kavya@gmail.com

Santhosh Thottingal

Swathanthra Malayalam Computing

santhosh.thottingal@gmail.com

1. Malayalam Unicode block: <https://Unicode.org/charts/PDF/U0D00.pdf>

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1.

Fluxus Editions, Brest, 2019, p. 329–351. <https://doi.org/10.36824/2018-graf-mano>

ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

	0D0	0D1	0D2	0D3	0D4	0D5	0D6	0D7
0	◌̇ 0D00	ഐ 0D10	ഠ 0D20	ഡ 0D30	റീ 0D40		ള 0D60	ഴ 0D70
1	◌̈́ 0D01		ഠ 0D21	ഡ 0D31	റു 0D41		ഴ 0D61	വ 0D71
2	ഠ 0D02	ഡ 0D12	ഢ 0D22	ണ 0D32	റു 0D42		ഴ 0D62	വ 0D72
3	ഠ 0D03	ഡ 0D13	ഢ 0D23	ണ 0D33	റു 0D43		ഴ 0D63	വ 0D73
4		ഠ 0D14	ഡ 0D24	ണ 0D34	റു 0D44	റ 0D54		ഴ 0D74
5	ഠ 0D05	ഡ 0D15	ഢ 0D25	ണ 0D35		ഴ 0D55		ഴ 0D75
6	ഠ 0D06	ഡ 0D16	ഢ 0D26	ണ 0D36	റ 0D46	റ 0D56	ഠ 0D66	ഡ 0D76
7	ഠ 0D07	ഡ 0D17	ഢ 0D27	ണ 0D37	റ 0D47	റ 0D57	ഠ 0D67	ഡ 0D77
8	ഠ 0D08	ഡ 0D18	ഢ 0D28	ണ 0D38	റ 0D48	റ 0D58	ഠ 0D68	ഡ 0D78
9	ഠ 0D09	ഡ 0D19	ഢ 0D29	ണ 0D39		ഴ 0D59	ഴ 0D69	ഴ 0D79
A	ഠ 0D0A	ഡ 0D1A	ഢ 0D2A	ണ 0D3A	റ 0D4A	റ 0D5A	ഠ 0D6A	ഡ 0D7A
B	ഠ 0D0B	ഡ 0D1B	ഢ 0D2B	ണ 0D3B	റ 0D4B	റ 0D5B	ഠ 0D6B	ഡ 0D7B
C	ഠ 0D0C	ഡ 0D1C	ഢ 0D2C	ണ 0D3C	റ 0D4C	റ 0D5C	ഠ 0D6C	ഡ 0D7C
D		ഠ 0D1D	ഡ 0D2D	ണ 0D3D	റ 0D4D	റ 0D5D	ഠ 0D6D	ഡ 0D7D
E	ഠ 0D0E	ഡ 0D1E	ഢ 0D2E	ണ 0D3E	റ 0D4E	റ 0D5E	ഠ 0D6E	ഡ 0D7E
F	ഠ 0D0F	ഡ 0D1F	ഢ 0D2F	ണ 0D3F	റ 0D4F	റ 0D5F	ഠ 0D6F	ഡ 0D7F

The Unicode Standard 12.1, Copyright © 1991-2019 Unicode, Inc. All rights reserved.

FIGURE 1. Unicode 12.1 Malayalam Code block

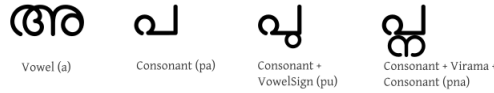


FIGURE 2. Few samples of graphemes in Malayalam

Malayalam script is abugida, or alphasyllabary. That is, consonant–vowel sequences are written as a unit: each unit is based on a consonant or conjunct letter, and vowel notation is secondary. Vowels are noted by modifying the consonants in the form of diacritics or vowel signs. Vowels have independent existence only at word beginnings. This is the common characteristic of the Brahmic family of from South and South-east Asia.

The script has acquired its uniqueness with its complex shaped ligatures formed by the consonants and conjuncts with signed vowel forms. Conjuncts are formed by a sequence of two more consonants. The conjunct grapheme usually has a shape smoothly blended from the constituent consonants. Figure 2 illustrates some samples.

2. Nature of Malayalam Script

Malayalam script as known today has 18 vowels, 39 consonants apart from various numerals, measuring units etc. Some are archaic in nature. The general nature of the script since its early days of evolution can be consolidated as below (Daniels and Bright, 1996; Varma, 2007).

1. Vowels and consonants are the basic building blocks of Malayalam script.
2. Vowels have stand alone existence in their pure form. Vowels in Malayalam are അ (a), ഇ (i), എ (e) etc. See Figure 3.
3. Vowels also appear as signed form modifying a consonant sound. Vowel signs have no existence without a consonant. Vowel signs in Malayalam are ി (i), ിഃ (i:), ൈ (e), ൈ (o)etc. See Figure 3.
4. Consonants in Malayalam always have the inherent vowel /a/, also known as *schwa* present in them. Consonants in Malayalam are ക (ka), ത (ta), ഗ (ga), ള (ḍa), ഡ (ḍ^ha) etc.
5. Any other vowel sound, other than /a/ associated with a consonant is written as a signed form of the consonant. The vowel signs can appear on the left, on the right or on both sides with respect to a consonant. Some signs modify the shape of base grapheme. Here are examples of consonants with vowel signs: കി (ki), ഗു (gu), ഡെ (ḍ^he), ളെ (ḍ^o). See Figure 4.

Independent Vowels in Malayalam Script															
അ	ആ	ഇ	ഈ	ഉ	ഊ	ഋ	ൠ	ഌ	ൡ	എ	ഐ	ഔ	ഓ	ഔ	അഃ
a	a:	i	i:	u	u:	ri	ri:	li	li:	e	e:	aj	o	o:	au

Dependant Vowel signs in Malayalam Script															
	ാ	ി	ീ	ു	ൂ	്രി	്രീ	്രി	്രീ	െ	േ	ൈ	ൊ	ോ	ൗ
a	a:	i	i:	u	u:	ri	ri:	li	li:	e	e:	aj	o	o:	au

FIGURE 3. The vowels in Malayalam. The independent vowels and dependant vowel sign forms are indicated along with their IPA in the bottom rows.

Consonant ക(ka) with various vowel signs															
ക	കാ	കി	കീ	കു	കൂ	ക്രി	ക്രീ	ക്രി	ക്രീ	കെ	കേ	കൈ	കൊ	കോ	കൗ
ka	ka:	ki	ki:	ku	ku:	kri	kri:	kli	kli:	ke	ke:	kaj	ko	ko:	kau

FIGURE 4. The consonant ക (ka) is shown with all possible associated vowel signs with corresponding IPAs indicated in the bottom rows.

- The removal of inherent /a/ in a consonant is marked in the script by a special character *virama*. Here is an example of consonant with *virama* (◌̣) sign: ക് (k). *Virama* after a consonant not only removes inherent /a/ but also indicates that there is no vowel sound following it.
 - ക (ka) + ◌̣ (*virama*) → ക് (k)
- A conjunct is formed by a sequence of consonants connected using *virama*. Examples:
 - ക (ka) + ◌̣ (*virama*) + ത (ta) → ക്ത (kta)
 - ഗ (ga) + ◌̣ (*virama*) + ദ (da) → ഗ്ദ (gda)
 A conjunct can again connect to another consonant using *virama* and form a longer conjunct as in:
 - ഗ്ദ (gda) + ◌̣ (*virama*) + ള (ḏa) → ഗ്ദḏ (gdḏa)
- Every conjunct can be modified by a vowel sign forming a new ligature.
 - ഗ്ദ (gda) + ◌̣ (*virama*) + ള (ḏa) + ു (u) → ഗ്ദയ് (gdḏu)
- Some consonants involved in the formation of conjuncts have signed forms. Example:
 - ക (ka) + ◌̣ (*virama*) + ര (ra) → ക്ര (kra)
 - ക (ka) + ◌̣ (*virama*) + ല (la) → ക്ല (kla)
 - ക (ka) + ◌̣ (*virama*) + യ (ja) → ക്യ (kja)

ക (ka)+[˘](virama) + ള (va) → ക്കുവ (kva)

ര (ra)+[˘](virama) + ക (ka) → ക്ക (rka) : This special sign is named *dot reph*.

10. Certain consonants have a unique grapheme representation in their pure form named *chillu*. ക്ക (k), ള് (l), ണ് (ṅ) , റ് (ṛ), ൽ (l), റ് (r)

The above nature of the script makes the number of unique graphemes to exceed 1,200 (Peani, 1772). Attempts of script reformation that occurred during the later half of 20th century aimed at simplifying the script to bring down the number of graphemes. Detailed discussion on script reform will follow in a later section.

3. Script in Early Printing Era

The shapes of conjuncts, relative positioning of signs and their sizes have changed over time to match the needs of writing methods.

The first ever book in Malayalam script was printed in Rome, in 1772. Printing technology demanded casting of movable types in huge numbers. Even though there were less than a hundred basic characters, the orthographic style demanded separate types for conjuncts, and their signed vowel forms. Apart from vowels, some consonants too have signed notations, further increasing the number of types needed in the foundry.

The first printed book in Malayalam using movable types, സംക്ഷേപവെദാരത്നം (*Samkshepavedartham*) in 1772 had more than thousand unique types (Cheriyān, 2008). It is a catechism book in the question answer form written by Clement Pianius². Figure 5 shows pages from the book. As it is perceivable from the figure, the script is mostly rectangular. The types were made in Rome and that is also where the book was printed.

Printing in Malayalam started natively during the 1820s (ibid.). The first native type casting and printing was done by Benjamin Bailey, an Anglican missionary in 1829. His contributions as a typographer made the curvy style of the Malayalam orthography popular (Nair, 1986). Figure 6a, shows pages from The New Testament printed using the types designed by Benjamin Bailey, printed in 1829 (Cheriyān, 2008). The script continued to evolve by separating some vowel sign types (്, ൾ) from the consonant or conjunct grapheme. Still the richness of conjuncts and their signed forms were largely retained. This can be seen in the page samples of the second edition of ശബ്ദതാരാവലി (*Sabdatharavali*), in Figure 6b, a comprehensive Malayalam dictionary prepared by Sreekantheswaram Padmanabha Pillai and published in 1930.

2. https://archive.org/details/SamkshepaVedartham_201311/page/n5

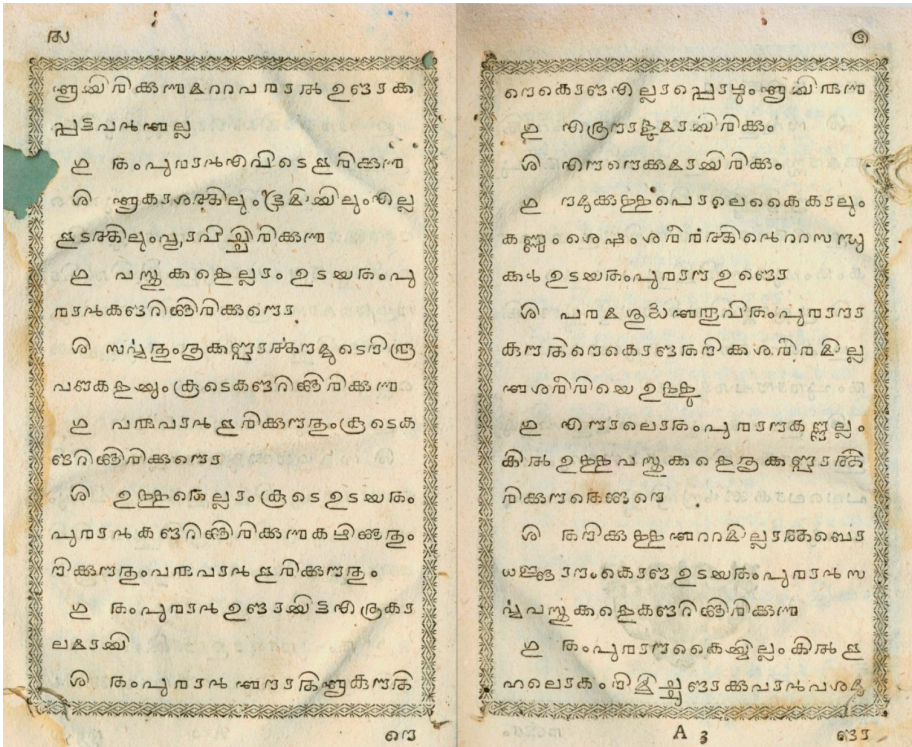
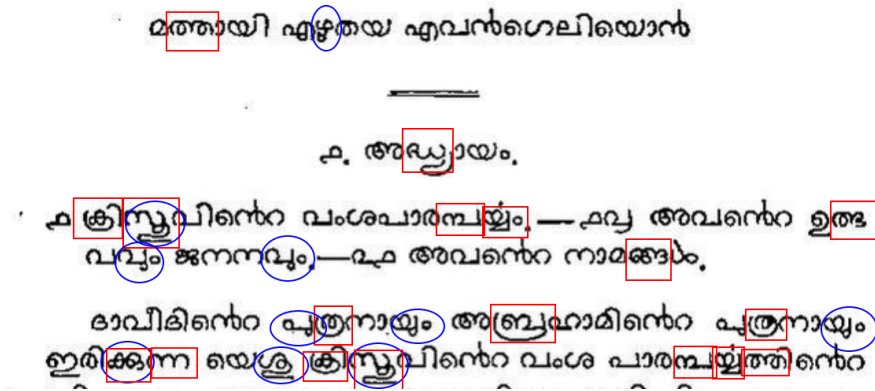


FIGURE 5. Pages from the book *Samksbepavedartham*, printed in 1772 at Rome. It is a catechism book written by Clement Pianius

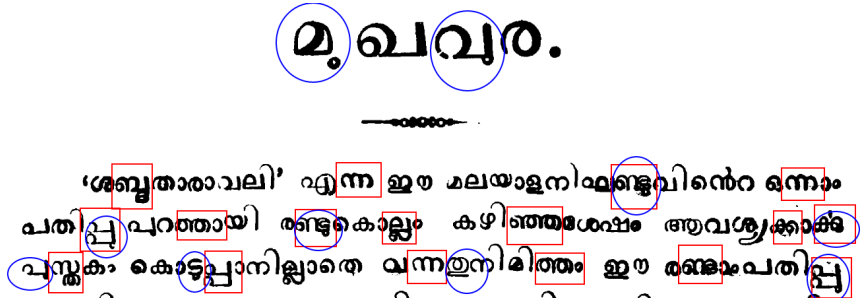
4. Script Reform

Typewriters became very popular around 1960s in Kerala. They used the design of English typewriters with the keys re-purposed for Malayalam. Obviously the keys were not enough to support all complex ligatures. The end result was Malayalam with all ligatures split up. It was a painful experience for reading and did not do any justice to the beauty of script as we can see from Figure 7.

To solve this problem, either the typewriter, or the language had to be redesigned. There were demands from newspaper and publishing industries to reduce the script complexity so that Malayalam becomes better suited for typewriters and printing. Based on this, in 1967 Kerala government appointed a committee to study script reformation. The committee submitted their report and in 1971 Kerala government published an order to reduce the complexity of the script (“Malayalam Script. Adoption of New Script for Use. Orders Issued” 1971).



(a) A page from New Testament in Malayalam published in 1829



(b) The Preface page of *Sabdatharavali*, a dictionary published in 1930

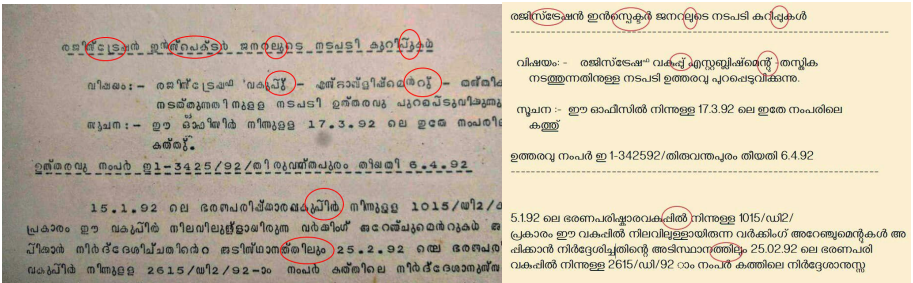
FIGURE 6. Samples of print documents. Complex graphemes formed by consonant sequences are indicated in rectangles and attached vowel sign forms are indicated in ellipses

The order was to discard the usage of complex conjuncts and to detach the vowel notations from the consonants and conjuncts. Being a forced intervention, this was a major event to be marked in the history of orthographic evolution.

The script variant of Malayalam that came into existence after the reformation order will henceforth be referred as *reformed orthography*. Reformed orthography consists only of a smaller set of graphemes than in the exhaustive set described in Section 2.



(a) Keyboard of Godrej typewriter repurposed for Malayalam



(b) On the left: Malayalam script typed by a typewriter. On the right: The same text rendered in a traditional orthography Unicode font

FIGURE 7. A typewriter and a sample of Malayalam document prepared using a typewriter. No complex graphemes are used in the document generated by the typewriter. Consonant sequences remain separated with virama (്) in between. The ellipses encircled in the left image of 7(b) represent how conjuncts are split up and vowel signs are separated. The corresponding rendering in a traditional orthography Unicode font is shown to the right in ellipses.



GOVERNMENT OF KERALA
Abstract

MALAYALAM SCRIPT-ADOPTION OF NEW SCRIPT FOR USE-ORDERS ISSUED

EDUCATION 'P' DEPARTMENT

G. O. (P) 37/71/Edn.

Dated, Trivandrum, 23rd March 1971.

Read: G.O. (P) 329/68/Edn.dated 11-7-1968

ORDER

The question of reducing the unwieldy number of alphabets and signs in Malayalam which consume much time and labour in the process of printing and typewriting, has been under consideration of Government for some time. In 1967 Government appointed a Committee with Shri Soornad P. N. Kunjan pillai, Editor, Malayalam Lexicon as convener to advise them on the question of reformation of Malayalam script. The committee in its report has made recommendations to reduce 75% of the total number of existing characters in printing and typewriting. The reformed Malayalam script recommended by the above Committee was revised with slight modifications by another committee appointed in 1969 to expedite the adoption of the new script for use. The recommendations of the above two committees in the matter of reformation of the Malayalam script are in brief as follows:

- i. ഉ, ഊ, ഋ, ഌ എന്നിവയുടെ മാത്രകൾ വ്യഞ്ജനങ്ങളിൽ നിന്നും വിടുവിക്കുക
- ii. പ്രചാരം കുറഞ്ഞ കൂട്ടക്ഷരങ്ങൾ ചന്ദ്രക്കല ഉപയോഗിച്ച് പിരിച്ച് എഴുതുക.

FIGURE 8. The Government Order on Malayalam Script Reform in 1971

Figure 8 shows the front page of government order proposing the new orthography style. The proposal aimed at reducing the grapheme usage in Malayalam by 75%. The major proposals of (“Malayalam Script. Adoption of New Script for Use. Orders Issued” 1971) are the following:

- Detach the signs of vowels ഉ (u), ഊ (u:) and ഋ (ri) from the base grapheme.
കു → ക്കു , ക്കു → ക്കു , ക്കു → ക്കു .
- Detach the consonant sign of ര (ra), that is റ , from the base grapheme
ക്ര (kra) → ക്ര (kra)
- Discard the usage of റ് in the consonant sequence in the form of diphthong sign . Instead use the alternate form റ . അക്കൻ → അർകൻ.
- Discard the use of rare conjuncts by splitting them down into constituent consonant sequence separated by the virama sign. Those re-

TABLE 1. Traditional and reformed orthography differences

No .	Characters	IPA	Traditional Orthography	Reformed Orthography
1	ക, റ	ka:ɾ	കറ	കാ
2	ദ, റ	ɖe	ദര	ദെ
3	ക, യ	kja	കയ	ക്യ
4	ക, ക	kka	കക	കാ
5	ക, ു	ku	കു	കൂ
6	ഗ, ു	gu	ഗു	ഗൂ
7	ഗ, റ	gɖa	ഗര	ഗാ
8	ഗ, ദ, ു	gɖu	ഗദു	ഗാദൂ
9	ഷ, ട	ʃta	ഷട	ഷാട
10	ക, റ	kra	കര	കാ
11	സ, ത, റ	sɖra	സതര	സാത്ര
12	ക, ര, ു	kru	കരു	കാറൂ

tained are: ക്ക, ഒ്ക, ഒ്ങ, ച്ച, ഞ്ച, ഞ്ഞ, ട്ട, ണ്ട, ണ്ണ, ത്ത, ന്ത, ന്ന, പ്പ, മ്മ, മ്മ, യ്യ, ല്ല, വ്വ. Others are split down as: ഗദ → ഗാദ .

- The signed form of consonants are to be separated from the base grapheme as in ക്യ, ക്ക, ക്ര .
- The signed *below base modifiers* of ല്ല (ല) may be retained as such പ്പല or split using *virama* sign as പ്പല .

As per the government order the reformed orthography would retain only 90 unique graphemes.

Table 1 compares the graphemes formed by sequence of basic characters in traditional and reformed orthography. As can be seen from the first four rows, the detached sign forms in traditional orthography are retained as such in the reformed one. Also some commonly used conjuncts are retained as such. The difference between two orthography variants becomes spectacular in the forthcoming rows. Complex ligatures formed by sequence of consonants gets split up by placing *virama* sign in between. Joined signed forms in traditional orthography get detached in the reformed variant.

It is important to note that the reformation order introduce the detached form of vowel signs for ഉ (u) and ഊ (u:) as ു and ൃ respectively. In the exhaustive set of traditional orthography u and u: had very diverse sign forms Manohar (2018). Their usage is evident from rows 5, 6, 8 and 12 in Table 1.

4.1. Adoption of Reformed Orthography

The print media switched to the reformed orthography to varying extents. The official prints of the government almost completely switched to the reformed style. Some publishers retained the graphemes for signed form of consonants but detached the signed vowel forms. Publishers adopted a set of conjuncts as per their choice and split down the others using *virama* sign. To quote from Daniels and Bright (1996):

However what happened is that individual printers opted for “modernizing” some characters but not others, thereby creating an inconsistent script with a large number of random options.

Students started to learn reformed orthographic style from the textbooks. But they continue to observe and learn the usage of traditional complex orthography widely seen in wall graffiti, poster designs and handwriting. Figures 9 and 10 illustrate the co-existence of both orthographies in the 1980s. The book in Fig. 9, is a text book published by the State Council for Educational Research and Training (SCERT)³, under the Government of Kerala, India for the third standard school students in 1988, and it uses reformed orthography. On the other hand in the Fig. 10 on can see a poster designed in traditional orthography for the popular movie നമുക്കുപാർക്കാൻ മുന്തിരിത്തോപ്പുകൾ [*Namukku Parkkan Munthirithoppukal*] “Vineyards for Us to Dwell In”⁴ directed by P. Padmarajan, produced by Ragam Movies, and released in 1986.

5. Script in Digital Era

The digitization of printing by the early 1990s was yet another remarkable event. The pre-Unicode digital fonts in Malayalam contained Malayalam glyphs mapped to the ASCII character space. Such fonts retained only a limited repertoire of conjuncts, because ASCII had a limitation of 128 code points (and other legacy fonts had similar limitations to 256 code points). Also the signed notations of vowels and consonants were detached from the base grapheme. Digital fonts before the Unicode era embraced the reformed orthography more closely. The publishing industry largely depended on these fonts for decades.

At the same time, people writing Malayalam in non-digital, non-printing contexts continued to use traditional orthography. Wall paintings, notice boards, artistic lettering used in magazines, movie titles continued using the traditional typography as illustrated in Fig. 11.

3. <http://www.scert.kerala.gov.in/>

4. https://en.wikipedia.org/wiki/Namukku_Parkkan_Munthirithoppukal

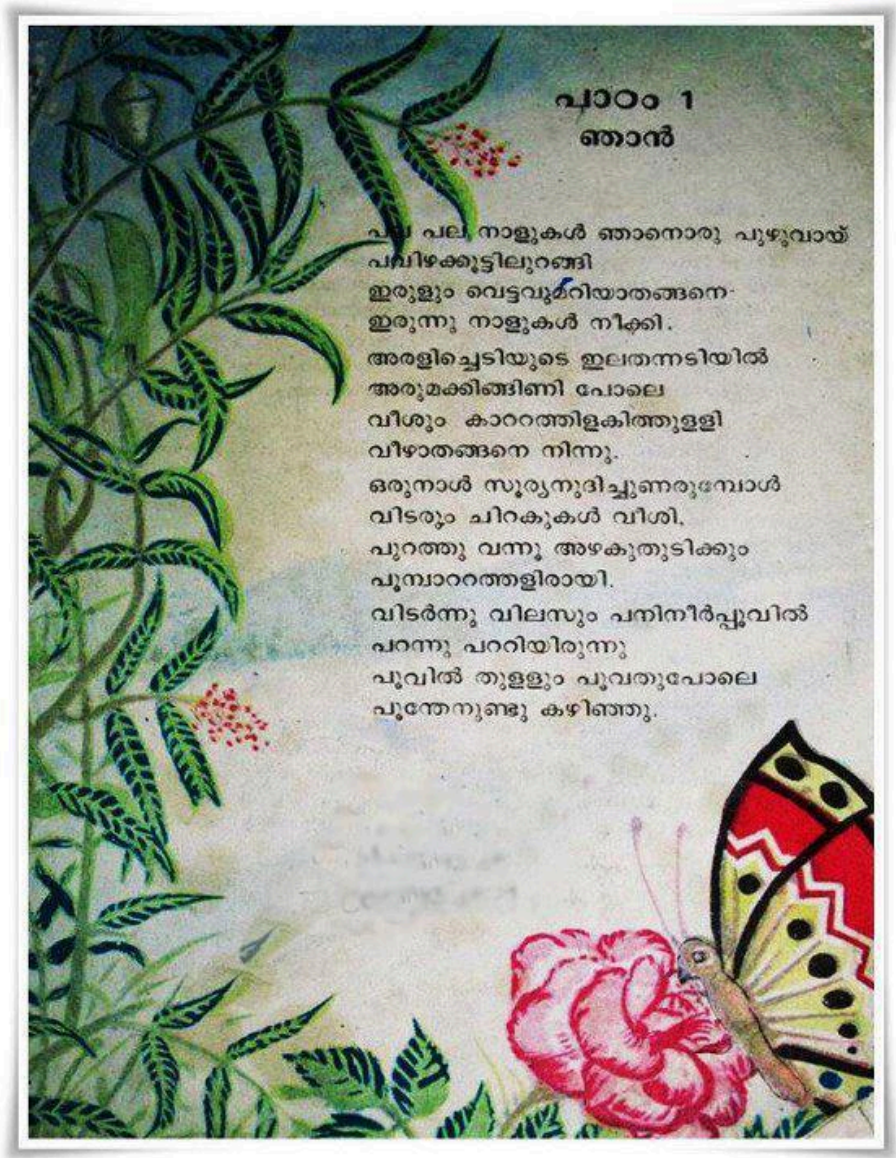


FIGURE 9. Reformed Orthography in a Malayalam textbook, published by the Government of Kerala in 1988

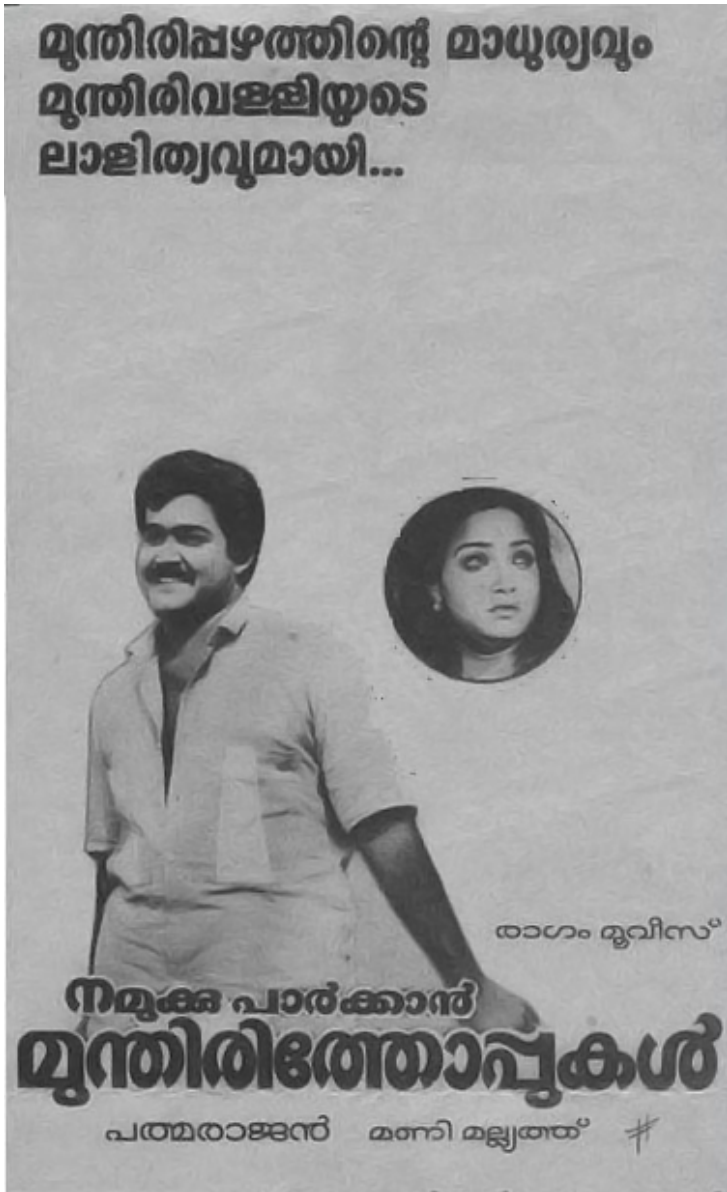


FIGURE 10. Poster of the Malayalam movie *Namukku Parkkan Muntbithoppukal*, “Vineyards for Us to Dwell In,” released in 1986, designed in traditional orthography



FIGURE 11. Sample of handwritten notice board by the Department of Forests, Government of Kerala. It is written in traditional orthography.

Rachana Aksharavedi, an organization formed in 1998, was successful in bringing back the exhaustive traditional orthography set with the help of then existing technology. At that time, Malayalam was not encoded in Unicode. *Rachana Aksharavedi* developed a font named *Rachana* with about 1,200 glyphs. Since it used no Unicode or OpenType technology, it was a set of 6 fonts, each covering about 200 glyphs mapped to 8-bit codepoints. A special editor known as *Rachana Editor* was required to automatically switch between these fonts and display Malayalam with data being English. This brave attempt was widely appreciated. A couple of years later, in 2001, Unicode encoded the Malayalam script.

5.1. Unicode and Advanced Digital Typography

Unicode did not differentiate between the traditional and reformed orthography. Orthography style was left to the typography and fonts layers. The ISO 639-1 standard for language names does not differentiate between the traditional and reformed orthography (Thottingal, 2016). This means that the digital representation using Unicode code points remains the same. Readers see that data using a font following traditional or reformed orthography as per their choice. Because of this abstraction,⁵ the reform had no impact on textual data processing.

5. *Note by the Editor.* This was unfortunately not the case for the 1982 Greek monotypic reform, which resulted in a loss of information so important that no Advanced Digital Typography technology will ever be able to reestablish.

TABLE 2. List of popular Unicode fonts in Malayalam

No.	Font	Orth. style	Vendor	Remarks
1.	Rachana	Traditional	SMC	Available in Ubuntu and other Linux distributions
2.	Meera	Traditional	SMC	"
3.	Manjari	Traditional	SMC	"
4.	AnjaliOldLipi	Traditional	SMC	"
5.	Chilanka	Traditional	SMC	"
6.	Dyuthi	Traditional	SMC	"
7.	Keraleeyam	Traditional	SMC	"
8.	Uroob	Traditional	SMC	"
9.	Manjari	Traditional	SMC	"
10.	Gayathri	Traditional	SMC	Produced by Kerala Bhasha Institute, Government of Kerala.
11.	Karumbi	Traditional	SMC	"
12.	Noto Malayalam	Reformed	SMC	"
13.	NotoSans Malayalam	Reformed	Google	Default in Android OS
14.	Nirmala	Reformed	Microsoft	Default in Windows OS
15.	Sangam	Reformed	Apple	Default in Apple products

With the advent of Unicode based digital typography, complex conjunct formations and their rendering were no longer an impossibility. With only the basic graphemes encoded in Unicode, any long sequence of consonants and signs could be mapped to a single conjunct grapheme in signed or unsigned form. Complex rendering rules of the script can easily be handled by modern rendering engines. With these technical advancements, fonts which could very well support the traditional orthographic scheme of the Malayalam script emerged.

The Rachana font was ported to Unicode. Parallel to that, more Unicode fonts emerged, notably AnjaliOldLipi. In 2006, Swathanthra Malayalam Computing (SMC), a free software developer community became active in Malayalam computing. Along with various language processing tools and technology improvements, SMC released a dozen of Malayalam fonts. With one exception, all fonts followed traditional orthography and embraced the OpenType technology.

GNU/Linux systems came with these traditional orthography fonts by default. Schools and government institutions were using GNU/Linux systems because of Kerala government policy to use Free Software. The user base of traditional orthography started to expand among digital

ഇവൾക്കു മാത്രമായ്

സുഗതകൃമാരി

ഇവൾക്കുമാത്രമായ്, കടലോളം കണ്ണീർ
 കുടിച്ചവൾ, ചിങ്ങവെയിലൊളി പോലെ
 ചിരിപ്പവൾ, ഉള്ളിൽ കൊടും തിയാളിടും
 ധരിത്രിയെപ്പോലെ തണുത്തിരുണ്ടവൾ.

ചവിട്ടാൻ, നിങ്ങൾക്കു ചിലപ്പോൾ പൂജിക്കാൻ,
 പരക്കെപ്പൂജിക്കാൻ, പരിത്യജിക്കുവാൻ,
 തുണയ്ക്കു കൈകോർത്തു നടക്കാൻ, മക്കളെ
 പിടയ്ക്കും നെഞ്ഞത്തു കിടത്തിപ്പോറ്റുവാൻ

ഇവൾക്കുമാത്രമായ് ഒരു ജന്മം; നെറ്റി
 തടഞ്ഞിലുണ്ടിവൾക്കൊരിറ്റു കൃങ്കമം,

വിളർത്ത ചുണ്ടത്തു നിലാച്ചിരി, ഹൃത്തിൻ
 വിളക്കുമാടത്തിലൊരു കെടാത്തിരി.

ഇവൾ ദൈവത്തിനും മുകളിൽ സ്നേഹത്തെ
 ഇരുത്തിപ്പൂജിപ്പോൾ, ഇവൾ കാലത്തിന്റെ
 കരങ്ങളിൽ മാത്രം സമാശ്വസിക്കുവോൾ.

ഇവൾക്കുമാത്രമായൊരു ഗാനം പാടാ-
 നെന്നിക്കു നിഷ്ഫലമൊരു മോഹം, സഖി....!

(സുഗതകൃമാരിയുടെ കവിതകൾ)

FIGURE 12. Malayalam Textbook – 2011. 8-bit based reformed orthography fonts in print. It was published by the State Council for Educational Research and Training (SCERT), Kerala, India for the tenth standard school students in 2011.

Malayalam users. The IT education curriculum in schools also widely used these fonts.

Table 2 lists the Malayalam Unicode fonts available by default in various operating systems. Availability of good quality traditional orthography fonts accelerated the usage of traditional orthography in digital space.

The typesetting tools and software adapted to 8-bit based fonts were the default in the publishing industry since 1990s. Even after the encoding of Malayalam in Unicode in 2001, the printing and publishing industry continued their practices. The book in Fig. 12 was published by the State Council for Educational Research and Training (SCERT)⁶, Kerala, India for the tenth standard school students in 2011. It shows the usage of ASCII based reformed orthography in school textbooks printed in 2011. This was largely due to lack of Unicode and complex script rendering support in major typesetting systems like Adobe InDesign. But these typesetting systems started supporting complex scripts and now we are seeing a highly accelerated adoption of Unicode and traditional orthography in print.

6. <http://www.scert.kerala.gov.in/>

6. Contemporary Script Usage

Currently everyone learns to read and write reformed orthography as part of school curriculum, but in fact is accustomed to the traditional orthographic style in everyday life. Non-digital media including wall writings, graffiti, bill-boards and handwriting continued using the traditional orthographic set. An example is shown in Fig. 13: a recent art installation in Sweetmeat Street, Kozhikode, Kerala, India, using traditional orthography.

Now that the technology has matured enough to support the traditional orthography, it is becoming more and more popular in the digital domain as well. There are newspapers and portals that switched to traditional orthography. Popular illustrated weeklies and science magazines switched to traditional orthography style in print (cf. Figures 14 and 15).

Figure 14 shows the popular illustrated weekly, സമകാലിക മലയാളം *Samakalika Malayalam* “Contemporary Malayalam,” announcing their return to traditional orthography in 2017 October stating the editorial decision summarized as:

Here returns the beauty of Letters: The weekly was always with the beautiful traditional orthography of Malayalam from the beginning. Later we had to stop it when the script reform occurred. But now that the traditional orthography is possible with the computers and with the fonts, we are returning to that orthography.

Kerala government is actively promoting Unicode usage in the official documents. Government orders are now mostly in Meera font, a traditional orthography font by SMC. Identity cards used for voting and public distribution system (See Figure 16) also uses traditional orthography. Gayathri, a traditional orthography typeface developed by SMC was sponsored by the Government of Kerala.

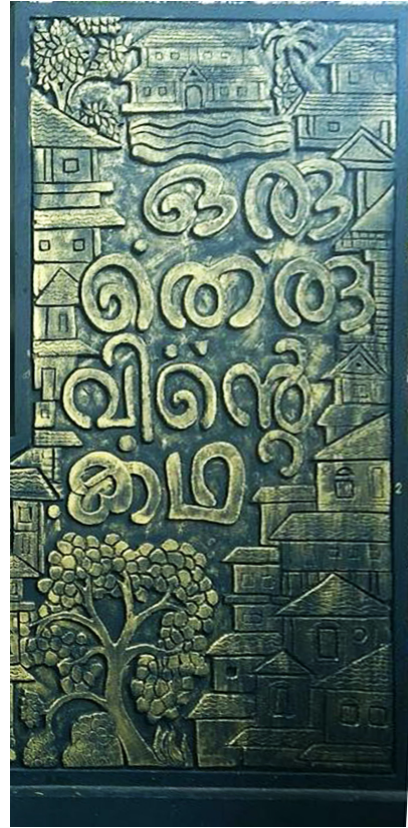


FIGURE 13. Art installation in Kozhikode, Kerala, India

The traditional orthography fonts by SMC are widely used in web content and social media memes. See an example in Figure 17.

Figure 18 shows the use of a mix of traditional and reformed orthography based title design for the movie തൺമുതലും ദൃക്സാക്ഷിയും *Thondimuthalum Driksakshiyum*, “The Mainour and the Witness,” directed by Dileesh Pothan, produced by Urvasi Theatres, and released in 2017.⁷ Another sample illustrating the usage of traditional orthography is in the title of the movie ഒരു മുത്തശ്ശി ന്റെ *Oru Muthassi Gadba*, “A Granny’s Mace,” in Fig. 19, directed by Jude Anthany Joseph, and released in 2016.⁸

7. Conclusion

Technology played a crucial role in defining the orthography of Malayalam from printing to digital age. When technology, such as typewriters, had limitations, the Malayalam script went through a difficult reform. But it flourished again with the help of digital technology. A single human generation has witnessed Malayalam’s transition from traditional orthography to reformed orthography and then again to traditional orthography. This fact has been illustrated through numerous examples in this paper.

Reformed orthography is sometimes referred as *modern* and traditional as *old*. But as traditional script is getting more popular in contemporary usage, calling it *old* may not be right. So we consciously avoided these terms in this paper.

References

- Cheriyān, Babu [ചരിയാൻ ബി.] (2008). ബഞ്ചമിൻ ബയെലിയും മലയാള സാഹിത്യവും [*Benjamin Bailey and Malayalam Literature*]. Kottayam: Mahatma Gandhi University, pp. 54–78.
- Daniels, Peter T. and William Bright (1996). *The World’s Writing Systems*. Oxford: Oxford University Press.
- “Malayalam Script. Adoption of New Script for Use. Orders Issued” (1971). Order by the Government of Kerala, India, G. O. (P) 37/71/Edn. <http://www.unicode.org/L2/L2008/08039-kerala-order.pdf>.
- Manohar, Kavya (2018). “u and ur: Vowel Signs of Malayalam”. <https://kavyamanohar.com/post/2018-04-15-u-vowel-signs-malayalam/>.

7. https://en.wikipedia.org/wiki/Thondimuthalum_Driksakshiyum

8. <https://www.imdb.com/title/tt5458448/>

- Nair, S. Gupthan [എസ്. ഗുപ്തൻ നായർ] (1986). ഗദ്യം പിന്തിട് വഴികൾ [*The Journey of Prose*]. Kottayam: DC Books, p. 108.
- Peani, Clemente (1772). *Alphabetum grandonico-malabaricum sive samscrudon-icum*. Rome: Stamperia della Sacra Congregazione de Propaganda Fide. https://numelyo.bm-lyon.fr/f_view/BML:BML_00G0001001370011032_19197.
- Thottingal, Santhosh (2016). "Proposal for Malayalam Language Subtags for Orthography Variants Rejected". <https://thottingal.in/blog/2016/09/30/malayalam-language-subtags>.
- Varma, A. R. Rajaraja [എ. ആർ. രാജരാജവർമ്മ] (2007). കരളപാണിനീയം [*A Grammar Book of Malayalam*]. 9th ed. Thrissur: DC Books.



തിരികെയെത്തുന്ന അക്ഷരലാവണ്യം

സ

മകാലിക മലയാളം വാരിക ഇരുപതു വർഷം പൂർത്തിയാക്കിയത് ഇക്കഴിഞ്ഞ രേയ് പതിനാറിനാണ്. തുടക്കം മുതൽക്കുതന്നെ അക്ഷര വിന്യാസത്തിലും രൂപകല്പനയിലും വാരിക പുതുക്കിയ്ക്കുന്നതിനായിട്ടുണ്ട്. അതിലൊന്നാണ്, പുതിയലിപി വിന്യാസത്തിൽ നഷ്ടപ്പെട്ടുപോയ ലാവണ്യവും പൂർണ്ണതയും തിരികെകൊണ്ടുവരാനുടനടത്തിയ ശ്രമങ്ങൾ. 'സാഹിത്യ വാരഫലം' പ്രസിദ്ധീകരിച്ചുകൊണ്ടിരിക്കെ, എം. കൃഷ്ണൻനായരുടെ ആഗ്രഹപ്രകാരവും നിർബന്ധത്തിലും എഴുത്ത്- വായനാ ജീവിതത്തിൽനിന്ന് അപ്രത്യക്ഷമായ പഴയലിപി ആ പാക്കിയിൽ ഉപയോഗിച്ചു. ലിപിസാങ്കേതികവിദ്യ വികസനമായിക്കൊണ്ടിരുന്ന അക്കാലത്ത് രചന എന്ന പേരിലുള്ള ലിപിയുടെ ഉപയോഗം ശ്രേഷ്ഠകരമായിരുന്നു. കൃഷ്ണൻനായർ സാറിന്റെ വിധേഹത്തെത്തുടർന്ന് പംക്തി അവസാനിച്ചതോടെ പഴയലിപി വാരികയുടെ താളുകളിൽനിന്നും അപ്രത്യക്ഷമായി. 1971-ൽ എൻ.വി. കൃഷ്ണവാര്യർ ടൈപ്പറൈറ്റുകൾക്കുവേണ്ടി രൂപകല്പന ചെയ്താണ് ഇന്നു നമ്മൾ ഉപയോഗിക്കുന്ന പുതിയലിപി. സാങ്കേതികപരാധ്യകളുടെ പശ്ചാത്തലത്തിൽ വൈകല്യങ്ങൾ പലതും പുതിയ ലിപിയുടെ ഭാഗമായി. എന്നാൽ മലയാളം യൂണികോഡിന്റെ കടന്നുവരവോടെ ഈ പരിമിതി മറികടക്കാമെന്ന ആശയം ഉരുത്തിരിഞ്ഞു. 'രചന' അക്ഷരലിപികളുടെ ശില്പി കെ.എച്ച്. ഇളയസെന്റ് ആത്മാർത്ഥത നിറഞ്ഞതും ആഴത്തിലുള്ളതുമായ അന്വേഷണമാണ് നമ്മുടെ അക്ഷരങ്ങളുടെ തനതുസൗന്ദര്യത്തെ തിരിച്ചുപിടിക്കാൻ ഉതകിയത്. സ്വതന്ത്ര മലയാളം കാമ്പ്യട്ടിങ് പ്രസ്ഥാനവും അതിന്റെ മുന്നണിപ്പോരാളികളായ നിരവധി ചെറുപ്പക്കാരാണ് ഈ നിശ്ശബ്ദ വിപ്ലവത്തിന്റെ വക്താക്കളും പ്രയോക്താക്കളും. ഇവർ അക്ഷരപ്രയത്നത്തിലൂടെ ആർജിച്ചത് 'സമകാലിക മലയാളം വാരിക' സ്വാംശീകരിക്കുകയാണ്. കെ.എച്ച്. ഇളയസെന്റ് പി.കെ. അശോക് കുമാറിനും നന്ദി. അക്ഷരത്തിന്റെ രൂപലാവണ്യവും പൂർണ്ണതയും പഴയലിപിയിലൂടെ തിരിച്ചെടുത്ത് വായനക്കാർക്ക് നൽകുകയാണ് വാരിക ഈ ലക്കം മുതൽ. ഇനി വായനക്കാർക്കുട്ട ഈ ലിപി വിപ്ലവത്തിന്റെ പ്രചാരകർ.



FIGURE 14. Samakalika Malayalam (Contemporary Malayalam), a popular illustrated weekly announcing their return to traditional orthography. 2017 October



സുരഭിവചന

ആഹ്വാനം... നമ്മുടെ



കേരളത്തിൽ കാണപ്പെടുന്ന ഏറ്റവും ചെറുതുള്ള പക്ഷികളിൽ ഒന്നാണ് വേഴാമ്പൽ. നിറം കൊണ്ടായാലും വലിപ്പം കൊണ്ടായാലും വേഴാമ്പൽ വേറിട്ടുനിൽക്കുന്നു. കൊക്കിന്റെ മുകൾഭാഗത്തുള്ള മുകൾമാണ് മറ്റു പക്ഷികളിൽ നിന്നും വേഴാമ്പലിനെ മാറ്റിനിർത്തുന്നത്. കേരളത്തിന്റെ സംസ്ഥാന പക്ഷി കൂടിയാണ് വേഴാമ്പൽ വിഭാഗത്തിൽപ്പെടുന്ന മലമുഴക്കി വേഴാമ്പൽ. ദേശീയ പക്ഷിനിരീക്ഷണദിനത്തിൽ നമുക്ക് മലമുഴക്കിയെ അൽപ്പമൊന്ന് അടുത്തറിയാം.

ആരാണീ മലമുഴക്കി?

ഇന്ത്യയിലെ ഏറ്റവും വലിയ വേഴാമ്പലിനും തന്നെ. മലമുഴക്കി വേഴാമ്പലിന് 3.9 കി. ഗ്രാം തൂക്കം കാണാം. പശ്ചിമഘട്ടത്തിലെ മലനിരകളിലും ഹിമാലയ താഴ്വാരങ്ങളിലും പ്രധാനമായും ഇവയെ കാണുന്നു. Great Hornbill, Great Pied Hornbill, Great Indian Hornbill എന്നീ ഇംഗ്ലീഷ് പേരുകളൊക്കെ മലമുഴക്കിയുടേതാണ്. കട്ടോടംചാത്തൻ, മരവിത്തലച്ചി, മരിത്ത

ലച്ചി എന്നൊക്കെ മലയാള പേരുകളും ഉണ്ട് ഇതിന്.



എന്താണ് 'മുഴ'ക്കുന്നത്?

മലമുഴക്കിയുടെ ചിറകുകളെ പരിചയപ്പെടാം. ഭിമാകാരമായ ഒന്നര മീറ്ററോളം നീളമുള്ള ഈ ചിറകുകൾ, ഇടയ്ക്കിടെ വിശ്രിയാണ് ഇവയുടെ പരക്കൽ. ചിറകിന്റെ പുറകുവശം നോക്കാം. അവിടെ, ചെറിയ തൂവലുകൾ മുളയ്ക്കുന്ന സ്ഥലം ഉള്ളോട്ട് വളഞ്ഞിരിക്കില്ല. അതി

FIGURE 15. Eureka, a famous science magazine for children changed to traditional orthography printing in November 2017.

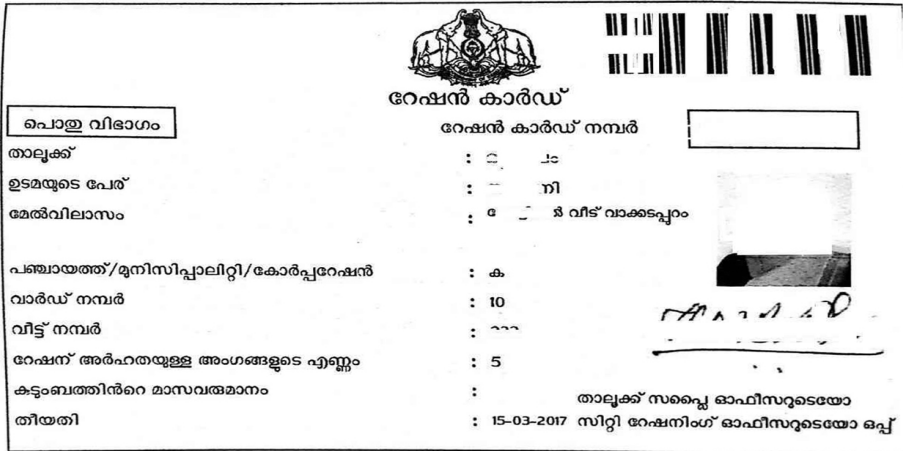


FIGURE 16. 2017 Identity card for public distribution system (Ration system). The Malayalam content is in traditional orthography font



FIGURE 17. Internet meme example in Malayalam from 2017 November, using traditional orthography font



FIGURE 18. Poster of the Malayalam movie, 'Thondimuthalum Driksakshiyum' (*The Mainour and the Witness*), 2017. The title uses a mix of reformed and traditional orthography.



FIGURE 19. Poster of the Malayalam movie 'Oru Muthassi Gadha' (*A Granny's Mace*), 2016. The title uses traditional orthography.


Investigating Keylogs as Time-Stamped Graphemics


Nicolas Ballier, Erin Pacquetet & Taylor Arnold


Abstract. This article investigates keystroke data, in an attempt to articulate the microlevel of the graphemic level with the macrolevel of text structures. Analyzing the time-stamps of keylogs, we suggest a hierarchy of constituents inspired by speech data and focus on the interaction of graphemic structure, phonological structure and textual structure within the dimension of time. We present the prototype of an R package designed to analyze keylog capture data, taking into account graphemic structures, syllable counts and parsing. Our R package under development offers functions that can be used to analyze the various levels of graphemic constituents produced by typists, from syllable counts to n -gram analysis.

1. Introduction

Current keyboards used with computers have reproduced mechanical and then electric keyboard layout (the QWERTY layout), even though alternative models such as BEPO or EWOPY (Bellis, 2017) have been developed now that the layout of keys on the keyboard is no longer dependent on the physical interactions of keys before hitting the ribbon. Typing is an emerging form of language production that has become part of our everyday lives in modern western societies. Most people use typing to write every day whether it is for professional or personal reasons. There is thus a need to better understand the processes involved in typing through a linguistic perspective, and it is interesting to consider

Nicolas Ballier  0000-0003-2179-1043
Université de Paris, CLILLAC-ARP, F-75013 Paris, France

Erin Pacquetet  0000-0001-9664-8167
Department of Linguistics, 609 Baldy Hall, University at Buffalo, North Campus, Buffalo, NY 14260-6420

Taylor Arnold  0000-0003-0576-0669
Department of Math & Computer Science, 212 Jepson Hall, 221 Richmond Way, University of Richmond, VA 23173

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1.
Fluxus Editions, Brest, 2019, p. 353–365. <https://doi.org/10.36824/2018-graf-ball>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

what linguistic information is encoded in typing. Moreover, typed language displays features from both traditional writing and oral speech, as well as features that are very specific to the typing medium and that do not have equivalents in other forms of language production.

Whenever a computer user types on a keyboard, it is possible to collect the timed typing information through keyloggers. The collection of user keystrokes on computers dates back to the beginning of personal computers and is still being used today for many of different purposes.

Keyloggers come both in hardware and software form (although nowadays, there are almost exclusively in software form) and are often devised as spywares that are hidden from the user's awareness. Historically, they have been used both by hackers wanting to recover passwords and by institutions trying to improve password and authentication security by learning individual typing patterns to discriminate between users (Giot, El-Abed, and Rosenberger, 2009), in particular to authenticate users in on-line courses.

Linguistically, keystroke logging is interesting because it enables researchers to witness the timed production of a typed text in a discreet and non-intrusive fashion and in a potentially naturalistic setting. Moreover, this technique requires easily accessible equipment (a computer and a keyboard). This is thus a very practical and accessible way of recording language production. Keystroke logging thus opens the door to not only investigate what language is produced but, and most importantly, how complex linguistic units are constructed and what the underlying processes are (Cislaru and Olive, 2018).

This article presents research in the making on the constituents of keylog capture data. The time stamps of the keys hit when we write texts have mostly been used to perform user authentication (Bergadano, Gunetti, and Picardi, 2002) and some datasets have been produced specifically for this aim, focusing on password typing (Giot, El-Abed, and Rosenberger, 2009; Giot, Ninassi, El-Abed, and Rosenberger, 2012). The past few years have seen an increase in the use of keystroke logging techniques in many areas of academic research. More recently, some studies have begun to address the linguistic data per se, whether to question non-canonical data (Plank, 2016) or to analyze the accelerations in typing (Van Waes, Leijten, and Neuwirth 2006; Leijten and Van Waes 2013). Some constituents have been investigated, either at the word level (Weingarten, Nottbusch, and Will, 2004) or above the word (Chukharev-Khudilaynen 2014; Cislaru and Olive 2016; Cislaru and Olive 2017), for exemple in synchronous computer-mediated communication (Charoenchaikorn, 2019) or in note-taking tasks (Malekian et al., 2019).

In the next section, we analyze the keystroke logs of English-speaking typists writing short essays in examination conditions (Charles C. Tappert, Cha, Villani, and Zack, 2012). We aim to characterize the

flow of typed data in terms of the size of the constituents processed by the typist (*processing chunks*). The aim is to establish the thresholds (and maxima) of relevant pauses to identify the constituents of typed texts, based on the model of the analysis of the prosodic constituents for the prosodic hierarchy (Nespor and Vogel, 2007). We compare, depending on the pauses identified, the span of typed sequences and their number of syllables; we explore possible constraints on the number of syllables cognitively treated for each identified constituent.

2. Datasets

As keystroke logging was primarily developed for spying and hacking, many tools available for keystroke logging are actually spyware. This is of course not desirable for academic research for ethical reasons as the logging has to be confined to the task presented to the test-takers and should stop when the experiment is over. Moreover, most spyware focuses rather on capturing the text typed than the timestamps of typing for the goal is to steal information and not to analyze typing patterns. Therefore, the kind of data collected and the way it is presented when using spyware is not suited to academic research. For instance, most spyware will collect the keys pressed and the timestamp of the typing session, but not the timepresses of each individual keys which are useful when looking at a production from a linguistic point of view.

In order to collect data in a safer and more controlled environment, several keylogging software packages have been designed specifically for research purposes. Among them, we can cite Inputlog (Leijten and Van Waes, 2013) which has been devised specifically for collecting keystrokes in an academic environment. Inputlog is a local keylogger that works with the software Microsoft Word. Once launched, the software opens a word document in which test takers can type freely. The keystrokes and mouse movements are recorded and saved. The software also performs analyses on the data and has a replay tool to re-watch the production. Inputlog is thus very well suited for academic research, but presents some limitations in the required setup for data collection as it has to be performed locally and there is little control over the parameters when using the predefined metrics.

Other tools are devised as sorts of hybrid solutions to collect data online and in an invisible way, but confined to a learning platform. An example of such a keylogger is the Moodle plugin BioAuth devised by Vincent Monaco (Stewart, Monaco, Cha, and Charles C. Tappert, 2011). Moodle is an open source learning management system that many universities worldwide use for course management. The BioAuth plugin enables course administrators to record keystrokes from students who are answering online quizzes hosted on the platform. The collection is lim-

ited to quizzes and stops once the answer is submitted—this makes it safe for students to use. The data is then stored on the Moodle database and can be accessed by the course administrator but not by the student. Students are identified on the platform and each production can be traced back to their typist. The platform also allows to embed various media types within the quiz question, which means that a wide variety of tasks such as picture description or guided production can be performed by the student.

When it comes to academic research, there are a lot of different datasets available for keystroke logging. Since research on the matter is fairly recent and there are no standards for keyloggers and/or experiment protocols, each research question calls for a different dataset and many studies end up collecting their own data, tailored to their needs. Therefore, there is a real plurality in terms of what is available.

In general, we can separate keystroke logging datasets into two categories: long-input and short-input datasets. Long-input datasets are made of long text input, usually answers to a question of at least one sentence. Examples of such studies include work on identifying typists based on stylometry and keystroke features (keystrokes dynamics-based user authentication) (Stewart, Monaco, Cha, and Charles C. Tappert 2011; Monaco, Stewart, Cha, and Charles C Tappert 2013; Kang and Cho 2015). These datasets make it possible to carry out linguistic analysis of the different language units and their mutual interactions.

Short-input datasets typically display typing sequences of a word or less. The most popular types of short-input studies are password studies where researchers attempt to gather information on how a specific typist types a specific password and use machine learning algorithms or biometrics to identify the typist and thus increase protection of accounts and personal data (Giot, El-Abed, and Rosenberger 2009; Killourhy and Maxion 2009). There is however little to no linguistic interest to such datasets as it is made out of very little language and passwords are often constructed as random sequences of characters.

Several datasets have been recently made publicly available. We will show how typing skills can be assessed by copying tasks and we will detail some of the resulting datasets. We use a long text input dataset of college examination answers presented in Charles C. Tappert, Cha, Villani, and Zack (2012). The keystrokes were collected from “40 students of a spreadsheet modelling course in the business school of a four-year liberal arts college” (ibid.). Although the test was administered online, the students did meet in a desktop classroom for each session, providing a controlled environment for the experiment. Tests were taken on Dell keyboards and desktops and the test takers got the opportunity to train on these keyboards beforehand. The test takers were not aware that their keystrokes were being captured at the time of the test. This is therefore a relatively natural setting for keystroke collection. The students took

four online tests of 10 questions each, with a two-week interval between each test. In the dataset, each test taker was assigned a number. A number is also assigned to each session. We used the keystrokes of 38 users, from user 2 to user 43 (users 10, 16, 20 and 36 are not part of the original dataset because they failed to complete the examination). For each key that was pressed by a given user during a given session, the dataset provides timestamps corresponding to the time at which the key was pressed and the time at which it was released. In addition, we are also given the keyname and the JavaScript keycode of each typed key. Table 1 is a sample from the dataset.

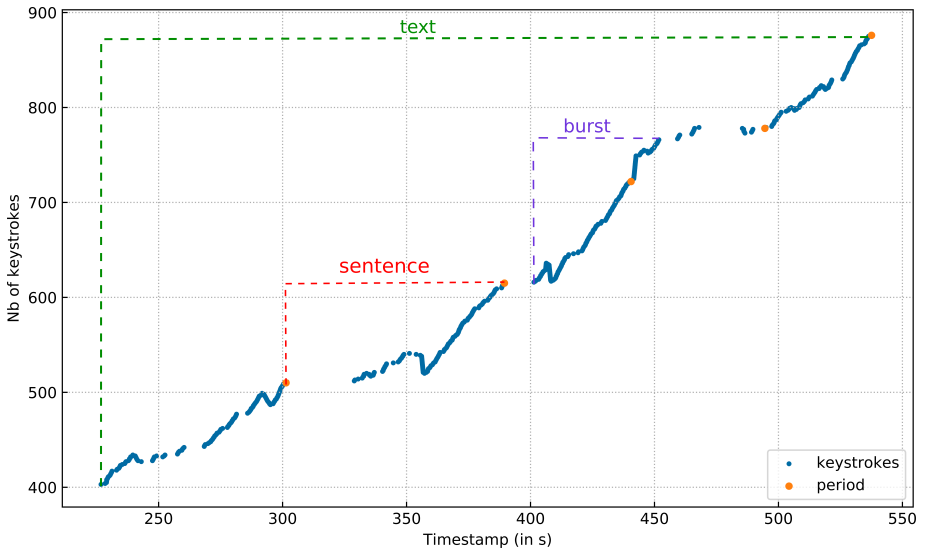
TABLE 1. Sample from the original dataset (Charles C. Tappert, Cha, Villani, and Zack, 2012)

user	session	timepress	timerelease	keycode	keyname
23	14	1301579856926	1301579857102	73	i
23	14	1301579857070	1301579857246	78	n
23	14	1301579857262	1301579857422	32	space
23	14	1301579858302	1301579858462	83	s
23	14	1301579858462	1301579858558	79	o
23	14	1301579858622	1301579858750	76	l
23	14	1301579858990	1301579859086	86	v
23	14	1301579859070	1301579859214	73	i
23	14	1301579859182	1301579859294	78	n
23	14	1301579859262	1301579859374	71	g
23	14	1301579859358	1301579859470	32	space

3. The Prosodic Hierarchy

Using our R package, we can reconstruct texts from this initial input and discuss the clustering of graphemes into higher constituents, whether at syllable, word or chunk level. Our research question can be summed up with Figure 1, which describes how chunks of graphemes (top) can cluster according to the prosodic hierarchy acknowledged in Nespor and Vogel (2007) (bottom) and whose lower constituents were tentatively described for keylogs (Weingarten, Nottbusch, and Will, 2004).

The *constituent model of written word production* (ibid.) distinguishes a graphemic word (W), some lexical constituents (LC) here aptly illustrated by the German compound *Flaschenöffner* (‘bottle opener’), syllables (S) and their phonological sub-constituents (O is the onset, R is for the rhyme), its graphemic layer (G_C stands for the consonant grapheme and G_{Cn} is a ‘consonant grapheme with n letters’ and G_V a vowel grapheme).



Prosodic domains	Syntactic units	Keyboard units
Speech	Text	Text
Paratone (\mathbb{T})	Paragraph	Paragraph
Phonological / prosodic utterance (PU)	(Utterance)	Chunk
Intonational Phrases (I or IP)	Sentence	
Phonological phrase (Φ or PP)	Clause Phrase Heavy NP	
Clitic group (C)	Noun Phrase	
Phonological / prosodic word (ω)		
Foot (Φ or F)	Word	Word
Syllable (σ)		
Mora (μ)		
Segments (phonemes)	Letter	Character

FIGURE 1. Mapping the series of bursts of a writer (top) to the hierarchical structure of the prosodic hierarchy (bottom)

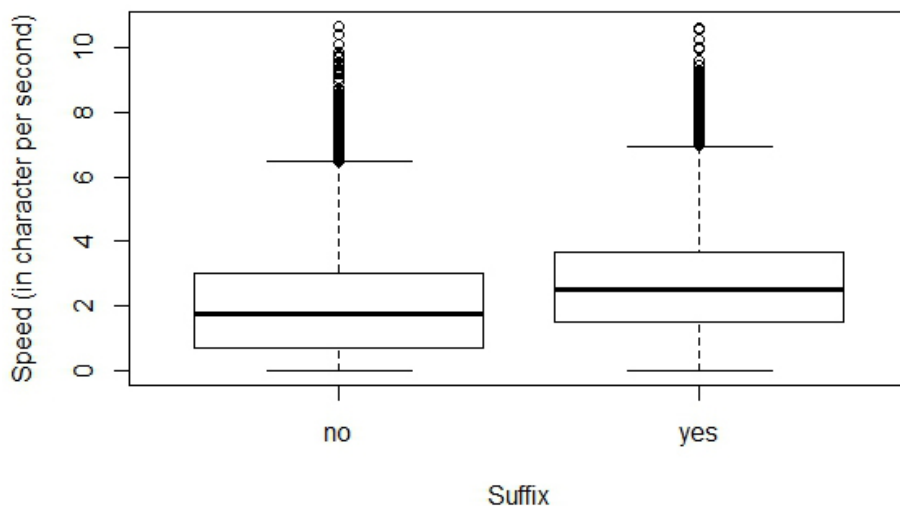


FIGURE 3. Boxplot of the speed of words with and without suffixes

TABLE 2. Frequency and length of words with and without suffixes

	No Suffix	Suffix
Average word frequency	21.09026	10.84965
Average number of characters	3.173415	5.853220

of writing and the complex interaction of revisions and corrections. The last section of the paper will show the benefits of our R scripts to compare the resulting texts and the dynamic processes of typing, especially the use of the backspace key. The dual nature of typed texts is summed up by Mahlow (2015) who advocated the need to address both “the product, i.e., the text where the error is visible for a reader, and the process, i.e., the editing operations causing this error.” We briefly illustrate graphs of inserted letters and repairs (backspace) and the resulting textual structures. As evidenced in the graphs below, we believe the ‘backspace’ key should be granted a special status it may erase complete textual bursts (right) so that we advocate a division of labour between ‘static’ and dynamic approaches of the keylogs.

5. Potential Applications for Learning Corpus Research

One of the main interests of using keystroke logging to analyze research production is that it allows researchers to collect and analyze data live.

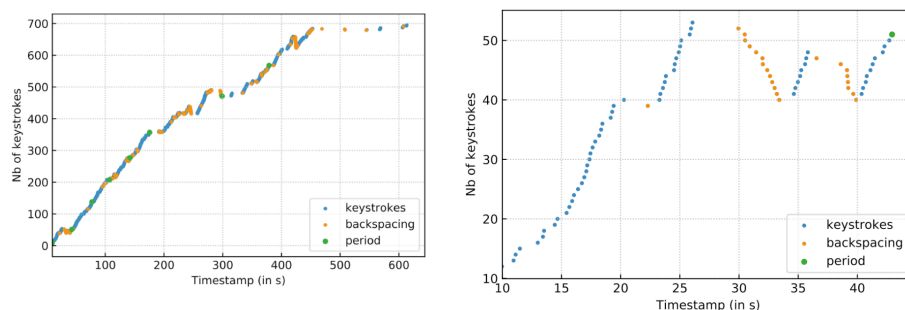


FIGURE 4. The potential complexity of backspacing and the need for a dynamic approach

This is particularly useful in an educational setting for it allows educators to provide visible feedback to students on the different aspects of their productions (Zhang, Zhu, Deane, and Guo, 2019).

Keystroke logging research presents interesting application possibilities, notably in language learning and teaching. When looking at the production of learners of a language in their target language, there are a few aspects that keystroke logging can inform us on that would not otherwise be accessible with only the final text as a resource. For instance, variables such as the amount of time spent on certain sections of the text or on difficult grammatical points are now available. It might also be interesting to look at how specific units such as reliability islands are produced. Editing, in the form of backspacing, is also made available by keystroke logging, which means that revision strategies are visible and can be analyzed.

When looking at keystroke data, it has been shown that four basic performance indicators were enough to separate typists into different clusters of learners that differed in writing processes and essay quality (*ibid.*). This could, in turn, lead teachers to better understand the needs of each specific student and to tailor their teaching to those needs.

Therefore, using keystroke features to investigate language production in an automated fashion will be useful to provide immediate and regular feedback to both students and educators.

This last section gives insight into learner data, perusing a portion of the data currently collected using Inputlog (Leijten and Van Waes, 2013) for the COREFL project (C. Lozano, A. Díaz-Negrillo, and Callies, to appear) at the university of Bremem to collect narratives. As can be seen in Fig. 5, writing bursts are not systematic.

In the second example (Fig. 6), we have manually represented the subdivisions of the writing task of a narrative based on a series of pic-

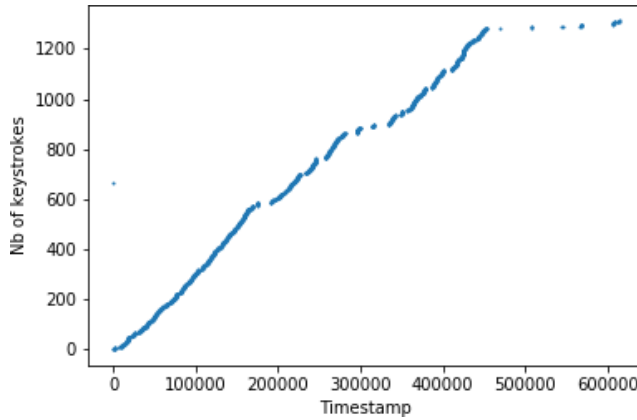


FIGURE 5. Visualization of typing bursts and picture changes in a narration task, data extracted from the COREFL corpus (A. M. C. Díaz-Negrillo and Cristóbal Lozano, 2018)

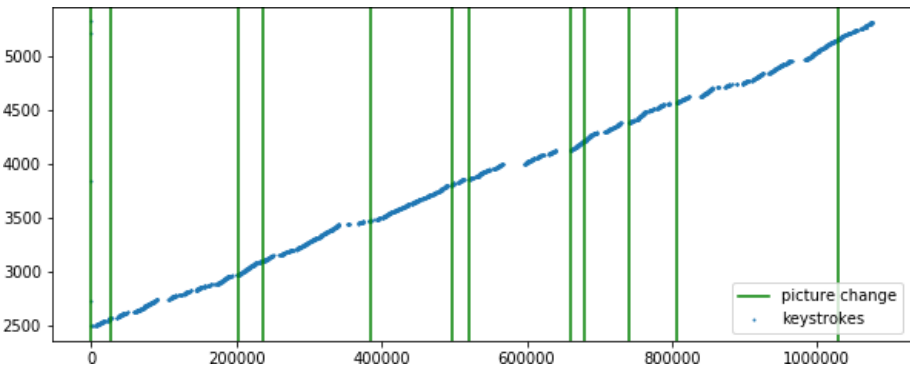


FIGURE 6. Visualization of typing bursts in time, data extracted from the COREFL corpus (A. M. C. Díaz-Negrillo and Cristóbal Lozano, 2018)

tures. In the figure, the vertical line represents a change from one picture to the other. As can be seen on the figure, some pictures require more description than others (as evidenced by the size of the window between each vertical green line), pauses may occur within one picture description, and the varying slopes correspond to different typing speeds for different pictures, which may in turn lead to question the difficulty of describing each picture.

6. Conclusion

In this chapter, we have suggested that the activity of writing with a keyboard shares features with speech in terms of potentially embedded constituents along a prosodic hierarchy. Our two-case studies with the two datasets considered allowed us to investigate only a fragment of the prosodic hierarchy. Whereas sublexical units such as suffixes have not seemed to be relevant, writing bursts and pauses call for investigations of units above the word such as collocations or reliability islands.

Analysing keystrokes gives an opportunity to reconsider Saussure's preference for speech over writing, as timepresses and time-release features act as features characterizing typed texts as time-stamped data, in a way similar to speech in spoken corpora. Aiming at analysing keylogs according to the prosodic hierarchy contextualises graphemes in relation to words, phrases, sentences and paragraphs, and therefore at text grammar level. It may not be the case that the variation of typing speed mirrors the variation of speech rhythm, but comparable grammars of chunking can be carried out for speech and keylog data.

References

- Bellis, Kouroch (2017). *La disposition Cœur 2.0 (ÉWOPY) comme disposition de clavier bureautique français: Réponse à l'enquête publique de l'AFNOR pour une norme PR NF Z71-300*. <https://hal.archives-ouvertes.fr/hal-01558613/document>.
- Bergadano, Francesco, Daniele Gunetti, and Claudia Picardi (2002). "User Authentication through Keystroke Dynamics". In: *ACM Transactions on Information and System Security (TISSEC)* 5.4, pp. 367–397.
- Charoenchaikorn, Vararin (2019). "L2 Revision and Post-task Anticipation during Text-Based Synchronous Computer-Mediated Communication (SCMC) Tasks". PhD Thesis. Lancaster University.
- Chukharev-Khudilaynen, Evgeny (2014). "Pauses in Spontaneous Written Communication: A Keystroke Logging Study". In: *Journal of Writing Research* 6.1, pp. 61–84.
- Cislaru, Georgeta and Thierry Olive (2016). "Les automatismes du scripteur: Jets textuels spontanés dans le processus de production écrite, le cas des constructions coordinatives". In: *SHS Web of Conferences*. Vol. 27. EDP Sciences, p. 06003.
- (2017). "Segments répétés, jets textuels et autres routines. Quel niveau de pré-construction?" In: *Corpus* 17, pp. 1–21.
- (2018). *Le processus de textualisation: analyse des unités linguistiques de performance écrite*. Louvain-la-Neuve, Paris: De Boeck Supérieur.

- Díaz-Negrillo, Ana Marcus Callies and Cristóbal Lozano (2018). "Designing and Compiling a Learner Corpus of Written and Spoken Narratives: The Corpus of English as a Foreign Language? (COREFL)". In: *ARISLA workshop (Anaphora Resolution in Second Language Acquisition)*. University of Granada.
- Evertz, Martin (in this volume). "The History of the Graphematic Foot in English and German".
- Giot, Romain, Mohamad El-Abed, and Christophe Rosenberger (2009). "Greyc Keystroke: A Benchmark for Keystroke Dynamics Biometric Systems". In: *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*. Washington, DC, pp. 1–6.
- Giot, Romain et al. (2012). "Analysis of the Acquisition Process for Keystroke Dynamics". In: *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)*. Darmstadt: IEEE, pp. 1–6.
- Kang, Pilsung and Sungzoon Cho (2015). "Keystroke Dynamics-Based User Authentication Using Long and Free Text Strings from Various Input Devices". In: *Information Sciences* 308, pp. 72–93.
- Killourhy, Kevin S. and Roy A. Moxon (2009). "Keystroke Dynamics—Benchmark Data Set". Carnegie-Mellon University, <http://www.cs.cmu.edu/~keystroke>.
- Leijten, Mariëlle and Luuk Van Waes (2013). "Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes". In: *Written Communication* 30.3, pp. 358–392.
- Lozano, C., A. Díaz-Negrillo, and M. Callies (to appear). "Designing and Compiling a Learner Corpus of Written and Spoken Narratives: COREFL". In: *What's in a Narrative? Variation in Story-Telling at the Interface between Language and Literacy*. Ed. by Christiane Bongartz and Jacopo Torregrossa.
- Mahlow, Cerstin (2015). "Learning from Errors: Systematic Analysis of Complex Writing Errors for Improving Writing Technology". In: *Text, Speech and Language Technology*. Vol. 48: *Language Production, Cognition, and the Lexicon*. Springer, pp. 419–438.
- Malekian, Donia et al. (2019). "Characterising Students Writing Processes Using Temporal Keystroke Analysis". In: *The 12th International Conference on Educational Data Mining*. Ed. by Michel Desmarais et al. Vol. 27. Montréal, pp. 354–359.
- Monaco, John V. et al. (2013). "Behavioral Biometric Verification of Student Identity in Online Course Assessment and Authentication of Authors in Literary Works". In: *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. Washington, DC, pp. 1–8.
- Nespor, Marina and Irene Vogel (2007). *Prosodic Phonology*. Vol. 28. de Gruyter.
- Plank, Barbara (2016). "Keystroke Dynamics as Signal for Shallow Syntactic Parsing". arXiv:1610.03321.

- Stewart, John C. et al. (2011). “An Investigation of Keystroke and Stylogometry Traits for Authenticating Online Test Takers”. In: *2011 International Joint Conference on Biometrics (IJCB)*. Washington, DC, pp. 1–7.
- Tappert, Charles C. et al. (2012). “A Keystroke Biometric System for Long-Text Input”. In: *Optimizing Information Security and Advancing Privacy Assurance: New Technologies*. Hershey, PA: IGI Global, pp. 32–57.
- Van Waes, Luuk, Mariëlle Leijten, and Christophe Neuwirth (2006). *Writing and Digital Media*. Leuven: Brill.
- Weingarten, Rüdiger, Guido Nottbusch, and Udo Will (2004). “Morphemes, Syllables, and Graphemes in Written Word Production”. In: *Trends in linguistics studies and monographs* 157, pp. 529–572.
- Zhang, Mo et al. (2019). “Identifying and Comparing Writing Process Patterns Using Keystroke Logs”. In: *Springer Proceedings in Mathematics & Statistics*. Vol. 265: *Quantitative Psychology*. Springer, pp. 367–381.

Vocalic and Consonantal Grapheme Classification through Spectral Decomposition


Patricia Thaine & Gerald Penn

Abstract. We consider two related problems in this paper. Given an undeciphered alphabetic writing system or mono-alphabetic cipher, determine: (1) which of its letters correspond to vowels and which to consonants; and (2) whether the writing system is a vocalic alphabet or an abjad. We are able to show that a very simple spectral decomposition based on character co-occurrences provides nearly perfect performance with respect to answering both question types.

1. Introduction

Most of the world's writing systems are based upon *alphabets*, in which each of the basic units of speech, called *phones*, receives its own representational unit or letter. The vast majority of phones are consonants or vowels, the former being produced through a partial or full obstruction of the vocal tract, the latter, through a stable interval of resonance at several characteristic frequencies called *formants*. In the course of deciphering an alphabet, one of the first important questions to answer is which of the letters correspond to vowels, and which to consonants, a problem that has been studied as far back as Ohaver (1933). Indeed, if there is disagreement as to whether a phonetic script is an alphabet or not, a near-perfect separation of its graphemes into consonantal and vocalic would be very important evidence for confirming the proposition that it was.

A well-publicized, recent attempt at classifying the letters of an undeciphered alphabet as either vocalic or consonantal was the one by Kim and Snyder (2013), who used a Bayesian approach to estimate an unobserved set of parameters that cause phonetic regularities among the distributions of letters in the alphabets of known/deciphered writing systems. By contrast, the method proposed in this paper is based on

Patricia Thaine · Gerald Penn  0000-0003-3553-8305
Department of Computer Science
University of Toronto
E-mail: {pthaine,gpenn}@cs.toronto.edu

Y. Haralambous (Ed.), *Graphemics in the 21st Century. Brest, June 13-15, 2018. Proceedings Grapholinguistics and Its Applications* (ISSN: 2534-5192), Vol. 1.
Fluxus Editions, Brest, 2019, p. 367–386. <https://doi.org/10.36824/2018-graf-thai>
ISBN: 978-2-9570549-0-9, e-ISBN: 978-2-9570549-1-6

a very simple spectral analysis of letter distributions within solely the writing system under investigation, and it requires no training or parameter tuning. It is furthermore based on a newly confirmed empirical universal over alphabetic writing systems that is interesting in its own right, and crucial to our method's numerical stability.

Spectral analysis of text for the purposes of vocalic/consonantal classification dates back to at least Moler and Morrison (1983), the method of which performs rather poorly. Our method can be regarded as both a simplification and improvement to Moler and Morrison. On average, our method correctly classifies 97.45% of characters in any alphabetic writing system.

Another notable antecedent is Goldsmith and Xanthos (2009), who discovered essentially the same method for vowel-consonant separation by spectrally analyzing phonemic transcriptions. While the premise that someone would have phonemically transcribed a text without knowing by the end which phones were vowels or consonants may seem far-fetched, Goldsmith and Xanthos (*ibid.*) draw some important conclusions for a subsequent analysis of vowel-harmonic processes that we shall not investigate further here. Goldsmith and Xanthos also cite Sukhotin (1962), whose method we evaluate below, as a precedent for their own study. Possibly, they were influenced in making this claim by Guy's (1991b) English gloss of Sukhotin's work, which misrepresents Sukhotin's (1962) intention as seeking to classify letters in a substitution cipher as vowels or consonants. Sukhotin's study, which was originally written in Russian, deals in fact with the written form (*bukv*) of plain text letters, and not of ciphers nor of the sounds of speech. Sukhotin begins his study by posing the research question of whether, given the well-known separation of the sounds of speech into vowels and consonants, there are similar classes for letters (*podobnyh klassah k'bukvam*). The distinction between written letters and phones is particularly salient in Russian, which, unlike English, has written letters that simply cannot be classified as vocalic or consonantal in any context or in isolation.¹

Sukhotin (*ibid.*) can be considered as an early attempt of our study of writing systems, but not of Goldsmith and Xanthos's (2009) study of phoneme clustering. In the present paper, we consider two applications of our method to the problem of classifying an alphabetic writing system as either an abjad (one with letters only for consonants) or a vocalic alphabet (one with letters for vowels as well). We then conclude with two initial studies, one of how the method may assist in interpreting

1. These are the front and back "yer" that respectively mark the presence or absence of palatalization. Sukhotin (1962) knew about the special status of these letters, too; when his method classifies the "front yer" as a vocalic, he expresses some satisfaction because the "front yer" did represent a vowel at an earlier stage in Russian writing.

	<i>_ *h</i>	<i>t*e</i>	<i>h*_</i>	<i>_ *a</i>	<i>f*t</i>	<i>a*_</i>	<i>c*t</i>
t	1	0	0	0	0	1	0
h	0	1	0	0	0	0	0
e	0	0	1	0	0	0	0
f	0	0	0	1	0	0	0
c	0	0	0	1	0	0	0
a	0	0	0	0	1	0	1

TABLE 1. The binary matrix, A , for the string ‘the fat cat’. Viewed as an adjacency matrix, it represents a bipartite graph.

historical linguistic data, and one of how the method may shed light on the decipherment of texts such as the Voynich manuscript.

2. A Spectral Universal over Alphabets

A *p-frame* (Stubbs and Barth, 2003) is reminiscent of a trigram context, except for the fact that it considers a preceding and a succeeding context element, rather than two preceding elements. The string ‘the fat cat,’ for example, contains these, among other p-frames at the character level: ‘*_ *h*,’ ‘*t*e*,’ ‘*h*_*,’ ‘*_ *a*,’ where ‘*_*’ represents a space.

Given a sufficiently long corpus C , in the alphabet Ω , let A be the binary matrix of dimension $m \times n$, where n is the number of different letter types in Ω and m is the number of different p-frames that occur in C (see Table 1), in which $A_{ij} = 1$ iff letter i occurs in p-frame j in C .

Every m by n matrix A has a singular-value decomposition into $A = U\Sigma V^T$. Usually, we are interested in Σ , a diagonal matrix containing the *singular values* of A , but we will be more concerned here with the n by n matrix V , the columns of which, the *right singular vectors* of A , are eigenvectors of $A^T A$. V is also *orthonormal*, which means that the inner product of any two right singular vectors $v_i \cdot v_j$ is 0, unless $i = j$, in which case the inner product is 1 (Strang, 2005).

If the rows and columns of U , Σ and V are permuted so that the singular values of Σ appear in decreasing order, then the first two right singular vectors are the most important, in the sense that they provide the most information about A . Let x and y be these two vectors; they are columns of V , and so they are rows of V^T , as shown in Figure 1. Empirically, each x_i is proportional to both the frequency of the i -th letter in C and the frequencies of the p-frame contexts in which the i -th letter occurs. Again empirically, each y_i ends up being proportional to the number of contexts that the i -th letter shares with other letters.

Because V is orthonormal, $\sum_i x_i y_i = 0$. Since their sum is zero, for some of the letters $i \in \Omega^+$, $x_i y_i$ is positive, and for other $i \in \Omega^-$, $x_i y_i$ is

$$A = U\Sigma V^T = \begin{pmatrix} \dots\dots\dots \\ \dots\dots\dots \\ \dots\dots\dots \\ \dots\dots\dots \\ \dots\dots\dots \\ \dots\dots\dots \end{pmatrix} \begin{pmatrix} \dots\dots\dots \\ \dots\dots\dots \\ \dots\dots\dots \\ \dots\dots\dots \\ \dots\dots\dots \\ \dots\dots\dots \end{pmatrix} \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \\ y_1 & y_2 & y_3 & y_4 & y_5 & y_6 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}.$$

FIGURE 1. Singular Value Decomposition of A

negative. The spectral universal we have empirically determined is that these two subsets of Ω almost perfectly separate the vocalic and consonantal graphemes of the writing system utilized by C . A moment's reflection will confirm that the p-frame distributions of vocalic graphemes are probably very different from the p-frame distributions of consonantal graphemes (Sukhotin, 1962), but the best thing about this universal is its inherent numerical stability. Table 2 shows the sums over these two sets for 15 alphabetic writing systems, expanded to 12 decimal places.

Language	$ \sum x_{\text{voc}} \cdot y_{\text{voc}} $	$ \sum x_{\text{cons}} \cdot y_{\text{cons}} $
Danish	0.461778253515	0.461778253515
Dutch	0.478014338904	0.478014338904
English	0.484420669972	0.484420669972
Finnish	0.471723103373	0.471723103373
French	0.482759327181	0.482759327181
German	0.440663056154	0.440663056154
Greek	0.447065776857	0.447065776857
Hawaiian	0.432782088536	0.432782088536
Italian	0.467317672843	0.467317672843
Latin	0.4656326487	0.4656326487
Maltese	0.496082609138	0.496082609138
Portuguese	0.463359992637	0.463359992637
Russian	0.491165538014	0.491165538014
Spanish	0.478974310472	0.478974310472
Swedish	0.430570626024	0.430570626024

TABLE 2. Inner products of x and y (Figure 1) for 15 different writing systems, accurate to 12 places

This calculation presumes a foreknowledge of what the vocalic and consonantal graphemes are, but if we were to order all of the letters in Ω by their value y_i , define a separator $y = b$, and then vary the parameter b so as to maximize the sum $|\sum_{i:y_i > b} x_i y_i| + |\sum_{i:y_i \leq b} x_i y_i|$, then $b = 0$ would

attain the maximum value. This is again trivial to prove in theory, but because the differences between vocalic and consonantal p-frames are the most important differences among all of the possible separators, we may observe empirically that $y = 0$ separates the vocalic graphemes from the consonantal ones. In other words, the actual values that the y_i attain are irrelevant; all that matters is their signs.

None of this provides any guidance as to which subset/sign contains the vocalic graphemes and which, the consonantal. Borrowing from the general idea behind Sukhotin's algorithm (Guy, 1991b), we will assume that the most frequent letter of any alphabet is vocalic² (Vietnamese is the singular exception that we have found to this rule), and thus label the subset that contains it as the vocalic container³. This yields Algorithm 1, which we evaluate in Table 3.⁴

3. Evaluating the Vocalic/Consonantal Identification Algorithm

Kim and Snyder (2013) report token-level accuracies with a macro-average of 98.85% across 503 alphabetic writing systems, with a standard deviation of about 2%. Token-level accuracies are somewhat misleading, as the hyperbolic distribution of letters in all naturally occurring alphabets makes it very easy to inflate accuracies even when the class of many (rare) letters cannot be determined. Furthermore, if the classified or readable portions of corpora were at issue, then these token accuracies should be micro-averaged, not macro-averaged, and, more importantly, they should be smoothed by an n -gram character model, to produce a more meaningful estimate.

Vocalic/consonantal classification is better viewed as a letter-type, not letter-instance, classification problem, in which progress is evaluated according to the percentage of letter types that are correctly classified. Semivocalic graphemes or whatever ambiguous classes one wishes to define should ideally be distinguished as extra classes, or at the very least disregarded. For a level comparison with our baselines (most are

2. Note that we treat <ò>, <ó>, <ô>, and <o>, for example, as four distinct graphemes.

3. Out of the 26 alphabets we examine, this assumption only fails for Vietnamese, whose most frequent letter is <n>. This is mainly due to the large number of diacriticized vocalic graphemes in Vietnamese that we treat individually.

4. In this and the subsequent experiments, the following writing systems were withheld as an evaluation set to prevent overfitting: Aramaic, Farsi, Hungarian, Serbian, Urdu, and Vietnamese.

All corpora were sampled from a combination of Wikipedia, Project Gutenberg and BBC World Service Web pages, and the sizes of texts vary between 14,316 and 706,422 characters (median=164,757). All punctuation was removed, and all letters were downcased.

Language	(Moler and Morrison, 1983)			Sukhotin's Algorithm			Algorithm 1			
	NC	P	R	A	P	R	A	P	R	A
Abkhaz	4	1.00	0.67	0.94	1.00	1.00	1.00	1.00	1.00	1.00
Afrikaans	18	0.71	0.36	0.31	0.93	0.81	0.88	1	0.81	0.91
Czech	23	1.00	0.63	0.68	1.00	0.94	0.98	1.00	0.94	0.98
Dutch	11	1.00	1.00	1.00	0.83	1.00	0.96	1.00	1.00	1.00
Danish	26	0.67	0.67	0.56	0.88	0.93	0.91	1.00	0.93	0.97
English (Middle)	4	1.00	1.00	1.00	1	0.90	0.96	1	0.90	0.96
English (Modern)	5	1.00	1.00	1.00	0.71	1.00	0.92	1.00	1.00	1.00
English (Old)	19	0.86	0.67	0.64	1.00	1.00	1.00	1.00	1.00	1.00
Finnish	3	1.00	0.89	0.96	0.89	1.00	0.96	0.89	1.00	0.96
French (Modern)	29	0.43	1.00	0.60	1.00	0.79	0.89	1.00	0.79	0.89
Inuktitut	6	1.00	1.00	1.00	0.95	0.95	0.95	1.00	0.95	0.97
Italian	17	0.90	0.90	0.86	0.91	0.67	0.82	1.00	0.93	0.97
German	13	1.00	0.88	0.93	0.73	1.00	0.89	0.88	1.00	0.96
Greek (Ancient)	3	0.83	1.00	0.95	1.00	1.00	1.00	1.00	1.00	1.00
Greek (Modern)	3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Hawaiian	5	0.90	0.90	0.92	0.83	0.91	0.90	1.00	1.00	1.00
Hungarian	14	0.44	0.80	0.71	0.94	0.94	0.94	1.00	1.00	1.00
Latin	3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Maltese	2	1.00	1.00	1.00	0.83	1.00	0.96	1.00	1.00	1.00
Portuguese	24	0.88	1.00	0.92	1.00	0.88	0.94	1.00	0.88	0.94
Russian	5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Spanish	25	0.89	0.89	0.85	0.90	0.69	0.88	1.00	0.82	0.95
Spanish	16	0.86	0.86	0.86	0.91	1.00	0.97	1.00	1.00	1.00
Swedish	6	1.00	1.00	1.00	0.89	1.00	0.96	0.80	1.00	0.93
Tagalog	4	1.00	0.94	0.97	0.95	1.00	0.97	1.00	0.89	0.95
Vietnamese	40	0.04	0.07	0.02	0.71	0.67	0.87	0.94	1.00	0.99

TABLE 3. Algorithm 1 evaluated with type-level accuracies. Corpora were sampled from the same sources as in Table 2, but with a number of characters between 25,738 and 968,298 characters (median = 177,529). The best accuracies are highlighted. Algorithm 1 incorrectly classifies several infrequent vocalic graphemes (<ë>, <ï>, <œ> and <ù> as consonantal) in Modern French. P, R, and A stand for Precision, Recall, and Accuracy, respectively. NC is the number of letters not classified by Moler and Morrison's (1983) algorithm; they are not necessarily semivocalic. Unclassified letters are not included in the calculation of Moler and Morrison's precision, recall, and accuracy, however; their results are even worse when NC letters are treated as false negatives.

Algorithm 1 Vocalic/consonantal classification algorithm

```

1:  $num_{words} \leftarrow 0$ 
2:  $num_{letters} \leftarrow length(letters)$ 
3:  $contexts \leftarrow list\ of\ num_{letters}\ empty\ lists$ 
4:  $frames_{keys} \leftarrow []$ 
5:  $frames_{values} \leftarrow []$ 
6:  $letters_{count} \leftarrow list\ of\ zeros\ of\ size\ num_{letters}$ 
7:  $A \leftarrow []$ 
8:  $A_{weighted} \leftarrow []$ 
9: function VOCALCONSClassification( $V, most\_freq\_letter$ )
10:    $coordinates \leftarrow zip(V[0], V[1], letters)$ 
11:    $cluster_1 \leftarrow triples\ where\ V[1]\ value\ >\ 0$ 
12:    $cluster_2 \leftarrow triples\ where\ V[1]\ value\ <\ 0$ 
13:    $voc \leftarrow cluster\ that\ has\ most\_freq\_letter$ 
14:    $cons \leftarrow cluster\ that\ does\ not\ have\ most\_freq\_letter$ 
15:   return  $voc, cons$ 
16: end function
17: function ALGORITHM1( $corpus, max$ )
18:   for all  $word \in corpus$  do
19:      $word \leftarrow ['\_'] + list(word) + ['\_']$ 
20:      $num_{words} += 1$ 
21:     if  $num_{words} > max$  then
22:       break
23:     end if
24:      $MakePFrames(word)$  # Calculates  $A$  and  $A_{weighted}$ 
25:   end for
26:    $index_{most\_freq\_letter} \leftarrow index\ of\ max(letters_{count})$ 
27:    $most\_freq\_letter \leftarrow letters[index_{most\_freq\_letter}]$ 
28:    $U, s, V \leftarrow SVD(A)$ 
29:    $voc, cons \leftarrow VocalConsClassification(V, most\_freq\_letter)$ 
30:   return  $voc, cons$ 
31: end function

```

interested in vocalic vs. non-vocalic; Kim and Snyder (2013) experimented with distinguishing nasals as well), ambiguous letters such as English ‘y’ have been manually identified and discarded altogether in Table 3.

It is impossible to determine the type accuracy of Kim and Snyder’s (ibid.) method, because they only made the raw counts of words in their corpus available⁵ (not the code, nor the resulting classifications). It is also impossible to reproduce their evaluation, since they did not provide their parameter settings. In addition, their ground truth classification of graphemes into vocalic and consonantal was remarkably ambitious. They treated all semivowels as consonantal, for example—even tokens where they act as vowels. The “front yer” palatalization marker in

5. http://pages.cs.wisc.edu/~ybkim/data/consonant_vowel_acl2013.tgz.

Russian Cyrillic was called consonantal, for example, and yet the “back yer” that blocks palatalization is called vocalic. With such arbitrary labellings of graphemes that simply should have been left out of the classification, a controlled comparison of even token accuracy is perhaps beside the point. For what it is worth, however, we could use the correct grapheme classifications in the 20 writing systems that constitute the overlap between the 503 that they sampled and the 26 that we did, and Algorithm 1’s macro-averaged token-accuracy on these is 99.93%, whereas Sukhotin’s is 96.05%.

An even greater cause for concern with this corpus is the sampling method that created it. Kim and Snyder’s (2013) use of a leave-one-out protocol to evaluate their method on each of their 503 writing systems at first seems reasonable—every known writing system should be pressed into the service of analyzing an unknown one. But all of these samples are Biblical, and many of them (the English, Portuguese, Italian and Spanish samples, for example, or the French and German samples) are the same verses translated into different languages. It is not reasonable in general to expect that a sample of unknown writing would necessarily be a translation of a text from a known writing system. The overlap in character contexts between transliterated proper names and cognates makes for a very charitable transfer of knowledge between writing systems.

Across the 26 writing systems that we have evaluated, our samples are all different texts from several genres. Our method requires no training, so all of the samples can be used for evaluation, but it also cannot avail itself of transfer across writing systems. On these samples, Algorithm 1 achieves a macro-averaged type accuracy of 97.45% and a macro-averaged token accuracy of 99.39% with a standard deviation of 1.67%. Performance is very robust in the realistic context of low transfer. On the same samples, Sukhotin’s algorithm has a macro-averaged type accuracy of 94.34%.

Moler and Morrison (1983)’s algorithm is less accurate than Algorithm 1. Moler and Morrison (*ibid.*) claim that their method is intended for “vowel-follows-consonant” (vfc) texts, where the proportion of vocalic graphemes following consonantal ones is greater than the proportion of vocalic following vocalic. Yet every writing system in our corpus is vfc, and still it performs poorly. Instead of using a binary adjacency matrix representing which letters occur within which p-frames, they calculate the number of times every possible letter pair occurs. They run SVD on the resulting matrix and use the second right and left singular vectors to plot the letters. The plot is divided into four quadrants, where letters in the fourth quadrant are classified as vocalic, those in the second quadrant as consonantal, and those in the first or third quadrants as “neuter,” [*sic*] meaning unclassified (see *NC* on Table 3). Our plots, on the other hand, are split into half planes with a crisp, numerically sta-

ble separation at the x -axis between the putative vocalics and putative consonantals, leaving no letter unclassified unless it falls on $y = 0$, which would only occur with completely unattested letters. Given the computational power and the number of electronic multilingual sources available at the time, Moler and Morrison (*ibid.*) had no workable means of thoroughly evaluating their method.

Another important concern is stability as a function of length—many undeciphered writing systems are not well attested in terms of the number or length of their surviving samples. Our spectral method performs robustly at the 97.45% level for sparse samples down to a minimum of about 500 word types or 4,000 word tokens. It is possible that below this threshold Sukhotin's algorithm would still be preferable.

Goldsmith and Xanthos (2009) only evaluate their method on one collection of written words, sampled from Finnish,⁶ and they obtain the same result as we do, misclassifying only the grapheme <q>.⁷ This should come as no surprise, because their method is an algebraically very close variant of ours—they compute eigenvectors on the Gram closure of our grapheme/context matrix (which they call F) instead of a singular value decomposition directly.

It may nevertheless come as a surprise that their method is so similar to ours. Their motivation consists of a lengthy discussion of graph cuts, along with a reference to Fiedler vectors, the name of the second eigenvector (the correlate to our y) of a graph's Laplacian matrix, which is known to relate to the graph's algebraic connectivity. Neither Goldsmith and Xanthos (*ibid.*) nor we explicitly calculate the Laplacian matrix of a graph, and if this would-be graph happened to have more than one connected component, the Fiedler vector would not be uniquely well-defined on its Laplacian matrix in general.⁸ Vocalic and consonantal graphemes rarely if ever separate into perfectly disjoint contexts; among our corpora the most disjoint is Vietnamese, in which vocalics and consonantals share exactly 100/645 p-frames. Out of curiosity, we evaluated our algorithm on the matrices from all 26 writing systems with their inter-CV/VC links removed. Performance degrades (macro-averaged accuracy: 89.08%)—which implies that this method is not merely computing an overall minimum graph cut—but not so badly

6. This is offered with the apology that Finnish is orthographically transparent, thus almost qualifying as a phonemic transcription.

7. Goldsmith and Xanthos's (2009) explanation for this is a "problem of threshold," but our study has found that the numerical stability of the threshold is extremely accurate. Instead, the problem is the relative disconnectedness of <q> from other graphemes owing to its sparsity, as the discussion in the next paragraph will elaborate upon.

8. Unless all of the connected components fortuitously had first and second eigenvalues of exactly the same magnitudes, the overall second non-zero eigenvector would not cross all of the components.

that partitions could merely be ignoring either all of the vocalics or all of the consonantals. The explanation found in Goldsmith and Xanthos (2009) therefore does not account for the robustness or generality of our collective approach. Our own determination of this method, along with this universal, was entirely experimental.

A final difference to our approach is that Goldsmith and Xanthos use bigram contexts instead of p-frames, although they are aware that this choice is arbitrary. Empirically, p-frames work better than bigrams (macro-averaged type accuracy: 89.06%) as well as trigrams with two preceding elements (96.24%).

Another pertinent study is that of Berg (2012), who evaluates his method only on English, German and Dutch orthography as well as a set of German phonemic transcripts. No quantitative measures are reported, but visual inspection of the figures provided is very reassuring. Berg used the entirety of morphologically preprocessed words as contexts, and used multidimensional scaling (MDS) rather than singular-value decomposition, so a precise comparison to ours is difficult.

Figures 2–7 shows example classifications by Algorithm 1 of six different writing systems. Each letter is plotted at its (x_i, y_i) coordinate, but the classification is made using only y_i . It is worth noting that semivocalic and other trouble-makers consistently fall very close to the $y = 0$ threshold. Maltese is particularly important, as it uses a vocalic alphabet with a Semitic language. Our correct handling of this case, and converse cases such as Farsi, demonstrates that we are responding to properties of alphabetic writing systems, and not of linguistic phylogeny.

4. Distinguishing Abjads from Vocalic Alphabets

Some writing systems assign syllabic or larger phonetic values to individual graphemes. Those that do not are sometimes called *alphabetic* writing systems, which is confusing because not all of them are true alphabets. There is another kind of alphabetic writing system called an *abjad*, which expresses only consonants. The Arabic writing system and other systems based on it (whether or not the underlying language is related to the Arabic language) are the prototypical abjads; the rest (e.g., Hebrew, Aramaic) expresses Hatto-Semitic languages. Abjads express words in languages that have vowels, but the vowels must be inferred from context, unless they are expressed through optional diacritics (Daniels and Bright, 1996).

We can use the spectral method presented in Section 2 to classify an alphabetic writing system as either an abjad or a true, vocalic alphabet. This is a different kind of classification problem than that of Section 3, as we are attempting here to classify the structure of entire writing systems rather than the phonetic values assigned to individual graphemes.

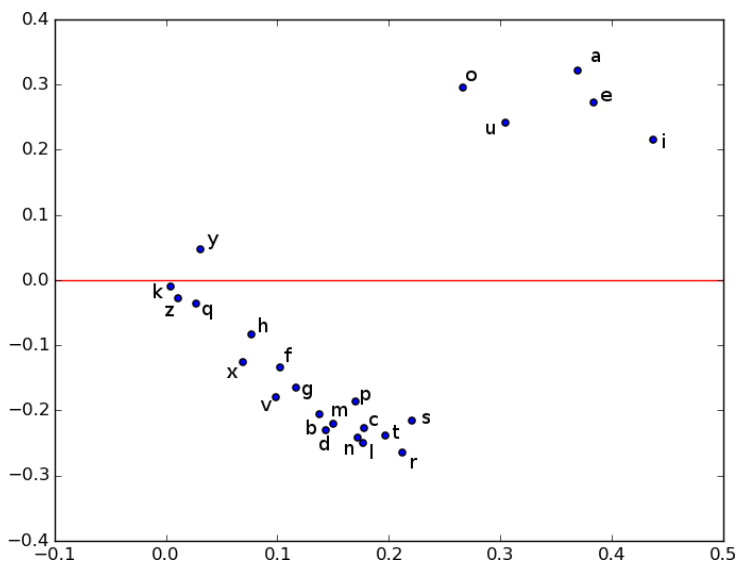


FIGURE 2. x and y for Latin

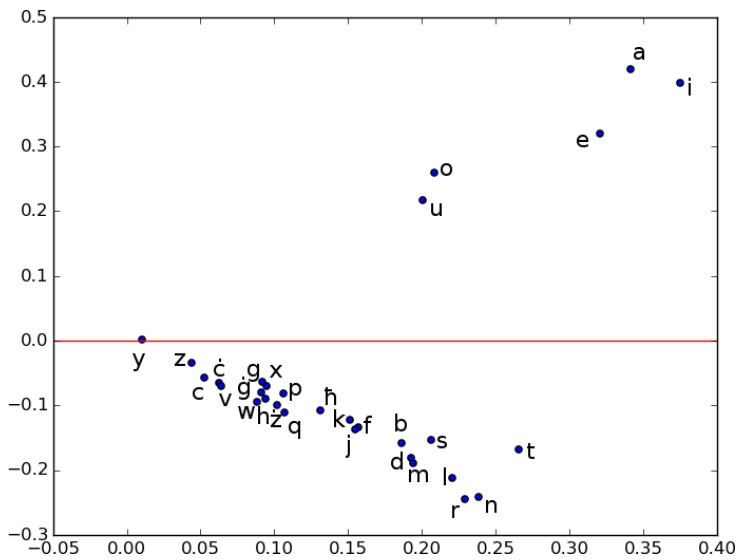


FIGURE 3. x and y for Maltese

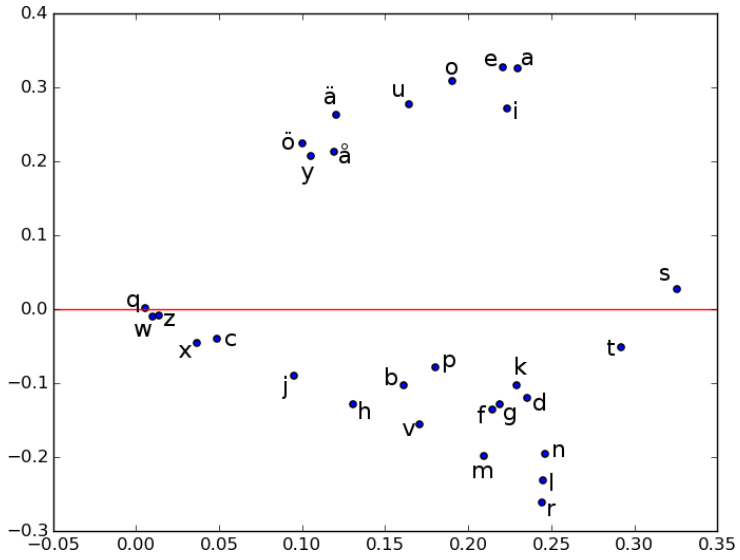


FIGURE 4. *x* and *y* for Swedish

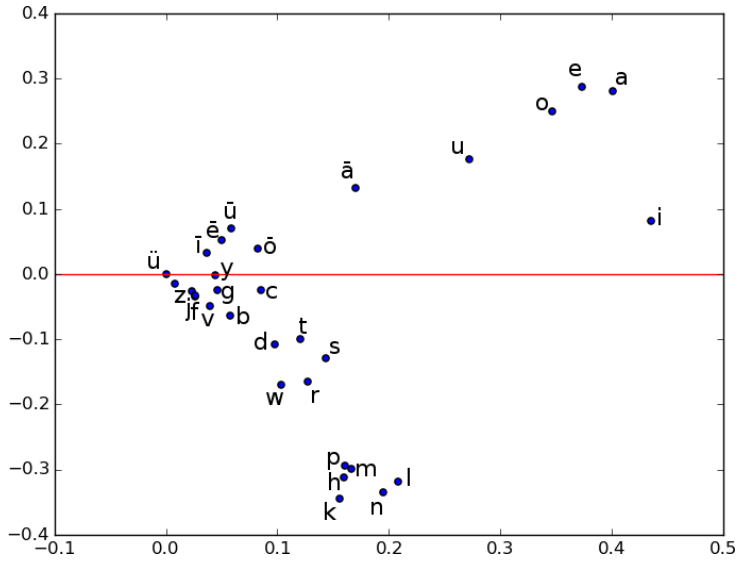


FIGURE 5. *x* and *y* for Hawaiian

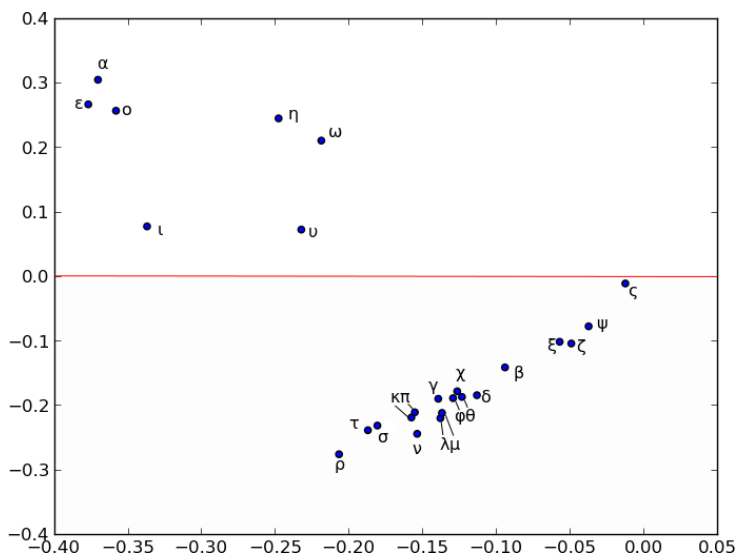


FIGURE 6. x and y for Modern Greek

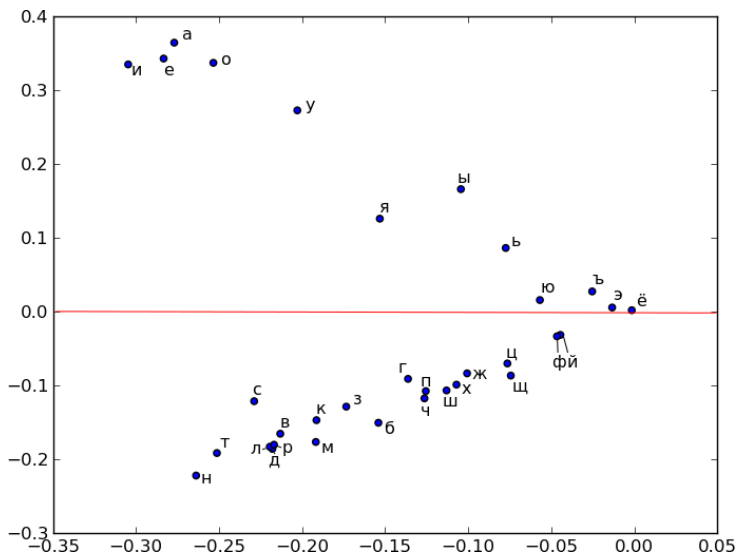


FIGURE 7. x and y for Russian

We will consider two algorithms for distinguishing abjads from vocalic alphabets:

4.1. Algorithm 2: Divergence

This variant begins by provisionally assuming that the writing system under investigation is a vocalic alphabet, and applying Algorithm 1 to it, which involves the calculation of the aforementioned matrix, \mathcal{A} , and the classification of every letter as consonantal or vocalic. There is a related matrix W , for which W_{ij} is the number of times letter i occurs in the context of p-frame j . W is not binary. We will label the rows of W as \hat{v}_i or \hat{c}_j according to whether i and j are labelled as vocalic or consonantal by Algorithm 1. Algorithm 1 still uses \mathcal{A} in assigning the labels, not W .

We can view each row of W as a discrete distribution over p-frame contexts. In recognition of this, Algorithm 2 calculates:

$$N = \sum_{\hat{v}_i, \hat{v}_j} |D|(\hat{v}_i \parallel \hat{v}_j) - \sum_{\hat{v}_i, \hat{c}_j} |D|(\hat{v}_i \parallel \hat{c}_j),$$

where $D(p \parallel q)$ is the Kullback-Leibler divergence of p and q . We use $|D|$ to represent the absolute-value of each element-wise calculation of $\hat{v}_i \log \frac{\hat{v}_i}{\hat{v}_j \text{ or } \hat{c}_j}$. The distributions of putative vocalics tend to be more dissimilar to one another in abjads than in true alphabets. The distributions of putative vocalics are more similar to that of putative consonantals in abjads than in true alphabets. Values of N are shown for 30 writing systems in Table 4. There, N separates the abjads from the vocalic alphabets at about $N = -100$.

4.2. Algorithm 3: Avocalic Words

For writing systems that conventionally use interword whitespace, we can alternatively apply vocalic grapheme identification to the task of discriminating abjads from vocalic alphabets by examining the percentage of word tokens with no vocalic graphemes.⁹ This method, Algorithm 3, is implicit to Reddy and Knight's (2011) 2-state HMM analysis of part of the Voynich manuscript, in which they observed that every word was recognized as an instance of the regular language a^*b . They argued that the most likely explanation is that every word was written

9. In vocalic writing systems, avocalic words include typographical errors, abbreviations and, in some writing systems, words with semivocalic graphemes that can occupy a syllabic mora, such as <y> in English.

Language	N	Language	N
Hungarian	773.7	Serbian	28.07
Tagalog	531.43	Modern Greek	20.6
Inuktitut	424.12	German	20.33
Vietnamese	359.53	French	16.01
Finnish	240.26	Modern English	-31.05
Old English	234.52	Portuguese	-53.19
Czech	223.96	Dutch	-57.18
Spanish	147.44	Afrikaans	-73.52
Russian	135.88	Italian	-89.94
Swedish	121.77	NVME	-167.63
Maltese	104.63	Farsi	-185.7
Latin	83.88	Aramaic	-191.23
Ancient Greek	65.88	Hebrew	-207.32
Hawaiian	57.29	Urdu	-220.01
Middle English	48.21	Arabic	-225.36

TABLE 4. Values of N for Algorithm 2, calculated over corpora of roughly 5,000 words each (min character tokens = 13,681, max = 39,936, median = 20,361). NVME is the Modern English corpus with vocalic graphemes removed. Abkhaz ($N = -70.94$) is not included because of its small size.

Language	V	C	Language	V	C
Arabic	3.75	0.92	Spanish	0.11	0.08
Hebrew	3.63	0.2	German	0.09	0.04
Urdu	2.58	0.22	Tagalog	0.07	0.06
Farsi	2.35	0.13	Inuktitut	0.07	0.05
Aramaic	1.97	0.18	Italian	0.07	0.04
NVME	0.19	0.69	Serbian	0.07	0.02
Abkhaz	0.63	0.44	Portuguese	0.05	0.05
Russian	0.37	0.29	Afrikaans	0.05	0.04
Maltese	0.36	0.06	Czech	0.05	0.01
Vietnamese	0.25	0.27	Modern English	0.05	0.01
Modern Greek	0.14	0.06	Latin	0.04	0.03
Dutch	0.13	0.04	Finnish	0.03	0.03
Old English	0.12	0.11	Swedish	0.03	0.03
Hawaiian	0.12	0.4	French	0.03	0.02
Middle English	0	0.12	Hungarian	0.02	0.01

TABLE 5. Percentages of word tokens with no putative vocalic (V) or consonantal (C) graphemes, as determined by Algorithm 3

with several consonantals followed by a vocalic, and that the Voynich manuscript therefore uses an abjad.

From this percentage, a decision boundary also emerges at about 1%, as shown in Table 5. NVME is not correctly classified unless one uses

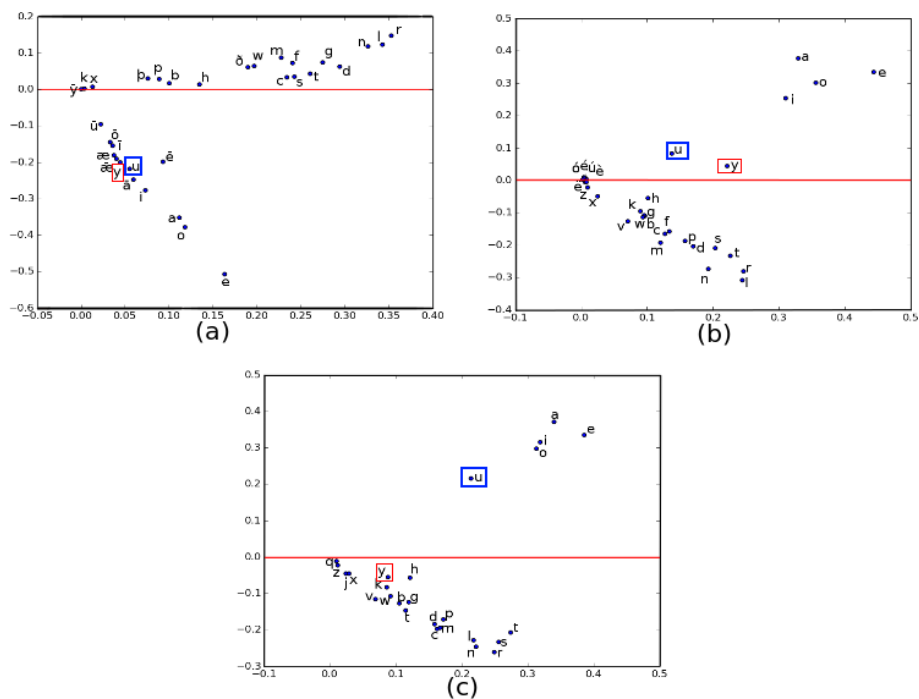
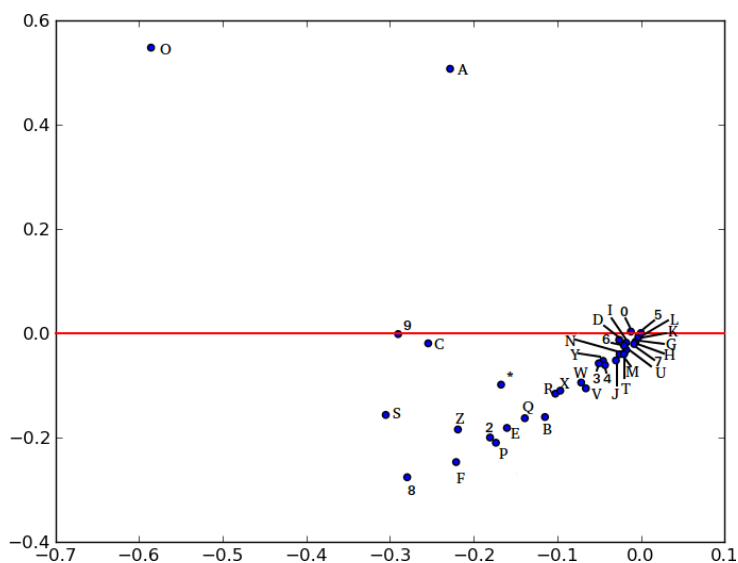


FIGURE 8. (a) Old English, (b) Middle English, (c) Modern English; Results shown are for texts approximately 25,700 characters long

the greater of the percentage of words without a vocalic or consonantal grapheme, but this (Modern English) with the once again putative vocalics and consonantals having been determined by Algorithm 1.

5. Change through Time

So far, we have applied Algorithm 1 to several different writing systems, treating them independently. Some writing traditions with long and well-documented histories, however, may present different spectral characteristics at different intervals along their documented timelines. Spectral decompositions of Old, Middle, and Modern English samples display evidence of several clear, well attested changes (Figure 8). For instance, it is readily apparent that a more dramatic modification of the writing system occurred between the Old English and Middle English periods than between Middle English and Modern English. Additionally, in Old English, <y> was used mainly as a vocalic grapheme. It became

FIGURE 9. x and y for Voynich A

more frequently consonantal with time. <u> became more vocalic in Modern English, because <u> and <v> had earlier been graphical variants of a single letter (Weiner, 2013).

6. The Voynich Manuscript

Guy (1991a) applied Sukhotin's method to two pages of the "biological" section of the Currier transliteration of the Voynich manuscript. The Currier transliteration uses typographical *, A-Z, and 0-9 in place of the cursive graphemes that appear in the manuscript in order to simplify its structural analysis. Currier had also found evidence for two separate writing systems within the manuscript, which he labelled "languages" A and B (Gillogly, 2002). The biological section was written primarily in language B. Guy (1991a) computed that <O>, <A>, <C>, and <G> are to be classified as vowels. Reddy and Knight (2011) state that "several" words in language B do not contain these characters, making it more likely that we are dealing with an abjad. Another possible conclusion would be that the Voynich manuscript is pseudo-writing, given its likely European provenance.

We have applied both Sukhotin's algorithm and Algorithm 1 to the entirety of both the sections identified by Currier as language A and, separately, the language B sections. The results for Algorithm 1 are dis-

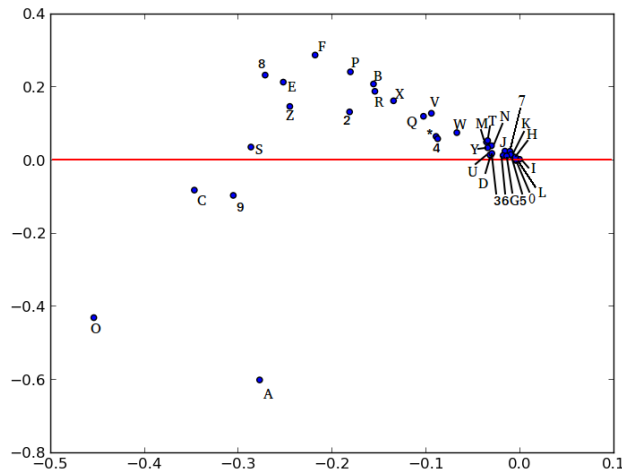


FIGURE 10. x and y for Voynich B

played in Figures 9 and 10. Although there is no ground truth with which to compare our results, Algorithm 1 outputs that $\langle A \rangle$, $\langle O \rangle$, and $\langle 0 \rangle$ are vocalic in language A. They do not occur in 5% of word tokens. Sukhotin's algorithm outputs $\langle O \rangle$, $\langle A \rangle$, $\langle 9 \rangle$, $\langle C \rangle$, $\langle 0 \rangle$, and $\langle 6 \rangle$ as the vocalic graphemes for language A, which do not occur in 0.77% of word tokens. Algorithm 1 and Sukhotin's Algorithm output the same vocalic graphemes for language B, namely $\langle C \rangle$, $\langle O \rangle$, $\langle A \rangle$, $\langle 9 \rangle$, $\langle L \rangle$, and $\langle 0 \rangle$. These do not occur in 0.53% of word tokens.

Algorithm 2 classifies both the A and B languages as vocalic alphabets, using Sukhotin's algorithm as the source for the putative vocalic/consonantal classification.

Given these results, we find it unlikely that either language A or language B is an abjad. It may even be the case that languages A and B have the same vocalic graphemes. The only vocalic grapheme posited by Sukhotin's Algorithm for language A but not for language B is $\langle 6 \rangle$ and the only vocalic grapheme posited for language B but not for language A is $\langle L \rangle$.

7. Conclusion and Future Work

We have shown that a very simple spectral decomposition based on character co-occurrences provides nearly perfect performance with re-

spect to classifying both a letter as vocalic or consonantal and a writing system as an abjad or alphabet. Algorithm 1 does not resolve other pertinent questions, e.g., distinguishing numbers from letters, or determining which capital letters correspond to which lowercase letters. Our method of vocalic/consonantal classification is meant to inform existing methods of finding graphemes' corresponding sounds. An additional source for associating sound values to graphemes is comparing letter frequencies between two related languages.

Future research on associating sound values to graphemes could include extending a method similar to Algorithm 1 to other types of writing systems, such as syllabaries.

References

- Berg, K. (2012). "Identifying Graphematic Units". In: *Written Language & Literacy* 15.1, pp. 26–45.
- Daniels, Peter T. and William Bright (1996). *The World's Writing Systems*. Oxford: Oxford University Press.
- Gillogly, Jim (2002). *Voynich Manuscript*.
- Goldsmith, J. and A. Xanthos (2009). "Learning Phonological Categories". In: *Language* 85.1, pp. 4–38.
- Guy, Jacques B.M. (1991a). "Statistical Properties of Two Folios of the Voynich Manuscript". In: *Cryptologia* 15.3, pp. 207–218.
- (1991b). "Vowel Identification: An Old (but Good) Algorithm". In: *Cryptologia* 15.3, pp. 258–262.
- Kim, Young-Bum and Benjamin Snyder (2013). "Unsupervised Consonant-Vowel Prediction over Hundreds of Languages". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, pp. 1527–1536.
- Moler, Cleve and Donald Morrison (1983). "Singular Value Analysis of Cryptograms". In: *American Mathematical Monthly* 90, pp. 78–87.
- Ohaver, Merle E. (1933). *Cryptogram Solving*. Columbus, OH: Etcetera Press.
- Reddy, Sravana and Kevin Knight (2011). "What We Know about the Voynich Manuscript". In: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Portland, OR, pp. 78–86.
- Strang, Gilbert (2005). *Linear Algebra and Its Applications*. 4th ed. Pacific Grove, CA: Brooks/Cole Publishing Company.
- Stubbs, Michael and Isabel Barth (2003). "Using Recurrent Phrases as Text-Type". In: *Functions of language* 10.1, pp. 61–104.

- Sukhotin, Boris V. [Сухотин, Борис В.] (1962). “Экспериментальное выделение классов букв с помощью электронной вычислительной машины [Experimental Selection of Letter Classes with the Help of Electronic Digital Machines]”. In: *Проблемы структурной лингвистики [Problems of Structural Linguistics]* 234, pp. 198–106.
- Weiner, Edmund (2013). *Early Modern English Pronunciation and Spelling*. <http://public.oed.com/aspects-of-english/english-in-time/early-modern-english-pronunciation-and-spelling/> [Accessed: 2014-07-29].

Index

Page numbers with suffix “de” refer to entries in German language text.
Entries in small caps refer to Unicode character names.

A

- Abbel, Éric, 68, 69
abbreviaturae, 20, 25
abjad, 17, 92, 146, 272, 367, 368,
376, 380, 381, 383–385
Abkhaz, 372, 381
abugida, 17, 146, 331
acrophony principle, 263
ADULT (emoji), 157
Advanced Digital Typography,
342
Afrikaans, 372, 381
AIDS, 274
Akkadian, 114
al-Khwārizmī, 3_{de}
Alekanō, 275
allograph, 230, 303
allophone, 129
alphabet, 17, 82, 92, 98, 99, 139,
146, 153, 174, 269, 271,
272, 275, 276, 282, 284–
287, 290, 291, 299, 317,
367–369, 371, 376, 380,
384, 385
alphabetic (noun), 139
alphabetical sorting, 304
alphabet
 phonemic, 272
alphagram, 145
alphasyllabary, 331
amant-es égaré-es, 68
ambisyllabicity, 35
Android, 23, 304
anime, 24
annotated text, 131
Arabic, 20, 21, 41, 42, 65, 106, 107,
115–117, 131, 142, 143, 147,
150–152, 154, 155, 245–
249, 376, 381
ARABIC LIGATURE BISMILLAH AR-
RAHMAN ARRAHEEM, 146
Aramaic, 247, 371, 376, 381
areal space, 130, 132
Armenian, 152
arroba, 54, 55
articulation
 double, 128, 153
 first, 129
 second, 129, 140
ASCII, 139, 158, 162, 168, 178, 339,
344
ateji, 191, 193
attachment point, 149
Austronesian, 275, 278

autonomistic approach, 129
 avocalic word, 380

B

backspace, 360
 backtrack, 46, 49, 70
barra, 69
 basic graphic unit, 214
 basic shape, 129, 139, 145, 151, 171,
 173, 177, 180
 BBC World , 371
 Belarusian, 98
 BEPO, 353
 Bible, 271, 296, 305, 374
 bidirectional algorithm, 150
binnen-I, 44, 58
 BioAuth, 355
 bit pattern, 139
 Brahmic, 331
 branching
 nucleus, 30, 31
 rhyme, 30
 breaking of the cursive line, 264
 Brest, 159
 Breton, 142
 burst, 358, 360–363

C

calendar, 4_{de}
 CANCEL TAG, 160
 case (grammatical), 84, 85
 case (letter), 58, 59, 98, 99, 103,
 146, 152, 371, 385
 inversion, 46, 48, 58, 81
 Catalan, 66
 catenation, 129
 operators, 148
 character, 18, 20–22, 24, 25, 47,
 66, 82, 92, 94, 96, 98, 99,
 104–106, 127, 132, 137–
 147, 149, 150, 152–158,
 160, 162, 167, 168, 170–
 174, 176–180, 190, 209–
 224, 228–242, 246, 262,

284–286, 288, 289, 293,
 294, 296, 299, 301, 303–
 305, 332, 333, 338, 339,
 356, 359, 360, 367–369,
 371, 372, 374, 381–384
 abstract, 139
 assigned, 191
 combining, 149
 encoded, 140
 identity, 144
 indicative, 228
 compound, 228
 simple, 228
 morpheme-representing, 193
 name, 144
 pictographic, 228
 polygraphic, 191
 semantic-phonetic, 228
 set, 24, 98, 103, 169, 180
 signific-phonetic, 228
 string, 127, 149, 151, 159
 two-component, 220
 unmotivated, 211
 with zero meaning, 220
 Cherokee, 99
 chillu, 333
 Chinese, 7_{de}, 11_{de}, 14_{de}, 21, 66, 107,
 132, 209–224
 dialects, 106
 numeral system, 3_{de}
 script reform, 13_{de}
chiocciola, 50
 classification
 vocalic/consonantal, 371
 Claudius (emperor), 19
 coda, 29, 30, 148
 code point, 141
 collocation, 363
 combination, 149
 interscript, 149
 combining
 character, 149
 character class, 151
 character sequence, 149
 common sense, 107

- commutation, 129
 component
 phonetic, 218, 235, 236
 semantic, 218
 signific, 235, 236
 pictographic, 236
 compositionality, 195, 196
 computer keyboard, 130
 Computer Modern, 138
 concatenation, 82
 CoNLL-U format, 131
 consonant, 19, 29, 31, 32, 35–38,
 150, 188, 246, 247, 263,
 293, 296, 298, 299, 301–
 303, 305, 311, 318, 319,
 329, 331–333, 335, 337–
 339, 343, 357, 359, 367,
 368, 374, 376
 ambisyllabic, 35, 38
 coronal, 298
 doubled, 35, 37
 geminated, 32–35
 intervocalic, 34
 labialized, 282
 nasal, 299
 prenasalized, 283
 South Arabian, 247
 constant meaning, 195
 constituent, 29, 30, 195, 209, 212,
 213, 216, 217, 228, 235,
 236, 241, 354, 355
 embedded, 363
 graph, 143
 higher, 357
 lower, 357
 model, 357, 359
 sub-, 359
 subsyllabic, 30
 unmotivated, 209, 211–215,
 219, 220, 223, 228
 contextual form, 154
 contour, 293
 control meaning, 139
coup de dés, 133
 Creve Cœur Police, 23, 176
 cuneiform, 120, 178
 Hittite, 121
 Mesopotamian, 246
 Sumerian, 19
 Currier transliteration, 383
 Cyrillic, 19, 69, 91, 92, 97, 99, 105–
 107, 141, 171, 178, 374
 Czech, 92, 142, 372, 381
 Côte d’Ivoire, 293–323
- ## D
- Da Vinci, 150
 Davis, Mark, 159, 160, 177, 180
 derivation, 148
 Devanagari, 106, 143
 diacritic, 50, 97–99, 102, 149, 259,
 269, 271, 272, 276–279,
 281, 282, 284–286, 288,
 289, 293, 296, 299, 300,
 304, 305, 310, 317–320,
 331, 371, 376
 density, 296, 318
 dot, 263
 primary, 263
 secondary, 260, 263
 stacked, 318
 stacking, 312
 dialect
 loyalty, 286
 diglossia, 113
 digraph, 44, 97, 113, 277, 278, 283,
 284, 286, 288, 291, 301,
 303
 digraphia, 111–123
 diachronic, 119
 diorthographia, 115
 distinctive stroke, 222
 double articulation, 128, 153
 Dtsch, 10_{de}
 dual nature
 of typed texts, 360
 Dutch, 31, 102, 152, 171, 370, 372,
 376, 381
 dvandva, 188

dwār letters, 263

dynamic underscore, 81

E

Eastern Dan, 293–323

écriture inclusive, *see* inclusive writing

écriture inclusive, 61

education, 361

element

 phonetic, 199

 semantic, 199

emicness, 93

emoji, 22–25, 128, 142, 146, 147, 155–162

EMOJI TAG BASE, 160

emoji

 prediction, 181

 sequence, 156

emoticon, 22, 23

English, 27–40, 42, 44, 51, 65, 92–94, 99, 102, 107, 111, 118, 129–131, 144, 169–172, 174, 188, 190, 192, 195–197, 210–212, 214, 232, 269–272, 275, 277, 281–283, 285, 287, 288, 291, 306, 334, 342, 354, 359, 368, 370, 372–374, 376, 380, 382

 American, 92

 British, 92

 Middle, 31, 381, 382

 Modern, 29–31, 35, 36, 381–383

 Old, 30, 31, 381, 382

Estonian, 102

estrangela, 247

eurocentrism, 7_{de}

EWOPY, 353

exhaustive tone marking, 294

exotype, 95

F

Facebook, 169, 172, 177, 181, 288

family resemblance, 105

Farsi, 371, 376, 381

♀-privileging, 47

Fibonacci, 3_{de}

FIELDATA, 139

Fijian, 288

finite automata, 82

finite state transducer, 148

Finnish, 188, 370, 372, 375, 381

first articulation, 129

flag sequence, 158

FoLiA format, 131

font, 18, 128, 133, 137, 138, 142–144, 149, 156, 157, 160, 162, 177, 336, 339, 342–346, 350

foot, 27–40, 296, 300–303, 314

 phonological, 31

formal language, 82

formant, 367

Fraktur, 103, 138

French, 37, 43–47, 50, 61, 72, 80, 81, 84, 86, 92, 99, 107, 115, 142, 149, 152, 260, 294, 303, 304, 307, 309–311, 315, 316, 370, 372, 374, 381

 Academy, 62

FULL STOP, 144

full tone marking, 294

G

gatekeeper, 144, 177

gemination

 graphematic, 34

 phonological, 34

gender, 41–86, 154–158, 174, 310

gender-neutral writing, 41–86

gender

 asymmetry, 47

 gap, 46

- replacement grapheme, 45
 separator grapheme, 46
 star, 46
 symmetry, 47
 general category, 146
 Georgian, 66
 German, 23, 44–47, 49, 50, 58, 65, 66, 81, 83, 84, 86, 99, 103, 111, 142, 144, 152, 153, 167, 169–172, 178, 228, 286, 357, 370, 372, 374, 376, 381
 Germanic runes, 178
 German
 modern, 35
 Old High, 30
 glottal stop, 285
 glottography, 174
 glyph, 18, 20, 43, 93, 99, 102, 103, 133, 137, 138, 143–145, 149, 150, 156, 158, 160, 177, 245, 246, 339, 342
 representative, 144, 145
 Go vote! Others do it too, 138
 grammatization, 260
 graph, 129, 137, 139, 142–144, 187
 graphematic
 foot, 27–40
 gemination, 32
 hierarchy, 28
 graphematics, 171
 suprasegmental, 27
 graphematic
 syllable, 28
 grapheme, 11_{de}, 35, 44–49, 51, 53–55, 57–59, 66, 68, 70, 72, 75, 81–84, 86, 91, 93, 99, 119, 127, 129, 133, 137, 139, 141–143, 145–150, 152–155, 170, 172, 175, 209–211, 214, 215, 220, 222, 224, 234, 235, 259, 260, 269, 271, 275, 277, 278, 283, 285, 289, 297–299, 302, 303, 305, 318, 329, 331, 333, 335–339, 343, 357, 359, 363, 367, 370–376, 380–385
 classification, 367–386
 dual nature, 153
 graphetics, 129, 136, 171
 perceptual, 136
 graphic
 meaning, 139
 graphicon, 168, 172, 181
 graphic
 unit, 218
 graph
 mathematical structure, 159
 grapholinguistics, 167
 Greek, 8_{de}, 24, 41, 44–47, 49, 50, 66, 73, 84, 86, 92, 93, 97, 113, 114, 117, 129, 131, 146, 248, 255, 259, 262, 342, 370, 379, 381
 Ancient, 372, 381
 Greeklish, 13
 Gutenberg galaxy, 21
- ## H
- hamza, 263
 hangul, 14_{de}, 148
 hanzi, 132, 145, 199, 200
 Hawaiian, 370, 372, 378, 381
 Hebrew, 24, 150, 257, 264, 376, 381
 Hegel, Georg Wilhelm Friedrich, 7–12_{de}
 hierarchy
 graphematic, 28
 phonological, 29
 hieroglyph, 7_{de}, 19, 118–121, 156
 Anatolian, 120, 121
 Hiri Motu, 275
 Hitler, Adolf, 138
 Hungarian, 371, 372, 381
 hybrid compound, 190
 hyphen, 304
 hyphenation, 132, 188

hyphen

soft, 132, 142, 177

hypothetical rectangle, 233

I

i malade, 303

Ibn Muqla, 262

iconicity, 229, 237

identification, 258

ideogram, 146, 173, 174, 210, 212,
218, 219

IJ digraph, 102

inclusive writing, 41, 41–86

indivisibility, 222

InKey, 277

Inputlog, 355

Instagram, 181

integrationist framework, 95

interlinear annotation character,
132

Inuktitut, 372, 381

INVISIBLE

SEPARATOR, 132

TIMES, 132

iOS, 23, 176

IPA, 10_{de}, 129, 285, 302, 332, 338

IRI, 137

ISO

10646, 22, 24

639, 342

693, 294

Basic Latin, 98, 102, 103

Italian, 44, 45, 48, 50, 83, 86, 111,
117, 121, 131, 142, 171, 370,
372, 374, 381

item (in Wikidata), 159

J

jamo, 148Japanese, 12_{de}, 14_{de}, 19–21, 24, 91,
92, 94, 104, 117, 132, 149,
155, 171, 180

K

kanji, 185–204

kanji-hiragana boundary, 190

kanji

unifying model, 203

Käsespätzle, 179keyboard, 23, 66, 130, 142, 150,
152, 157, 285, 286, 304,
353

keylogger, 354–356

Keyman, 277

keystroke feature, 356

Khmer, 147, 150

kokuji, 190, 200

Korean, 65, 117, 148, 191

Kufic calligraphy, 133

Kullback-Leibler divergence, 380

kun, 190, 191, 200, 201*kāishū*, 232, 233

L

Lafcadio Hearn, 12_{de}

Lao, 150

Latin (script), 91–107

Latin, 19–21, 24, 92, 115, 117, 141,
370, 372, 377, 381

LATIN SMALL LIGATURE IJ, 153

LATIN SMALL LIGATURE OE, 153

Latvian, 142

LEFT-TO-RIGHT MARK, 150

Leibniz, Gottfried Wilhelm, 8_{de}

letter distribution, 368

Levant, 248

LGBTQI communities, 81

Li-character, 235

Liberia, 294

ligature, 18, 20, 25, 127, 128, 157,
329, 331, 332, 334, 338

discretionary, 152

esthetic, 152

intermorphemic, 153

linguistically motivated, 152

mandatory, 151

- linear space, 132
lišbū, 232
 literal (in Wikidata), 159
 Lithuanian, 102
 logical order, 149
 logogram, 145, 146, 174, 181
 logograph, 192
- M
- macro-script, 106, 107
 macrographetic model, 135
 macrotypography, 135
 Malayalam, 329–347
 σ-privileging, 47
 Mallarmé, 133
 Maltese, 370, 372, 376, 377, 381
 MAN (emoji), 157
man'yōgana, 189
 Mao Zedong, 5_{de}
 MARK (method), 46
 marked nasalization, 287
 markup languages, 131
 mathematical notation, 129
 MathML, 132
 McLuhan, Marshall, 21
 media file (in Wikidata), 159
 Melpa, 275
 mesographetic model, 135
 mesotypography, 135
 middle dot, 65
 minimal graphic unit, 214
 mixed text, 25
 mnemonic component, 228
 modifier sequence, 156
 mono-alphabetic cipher, 367
 monomorphemic word, 192
 Moodle, 355
 morpheme, 57, 82, 129, 132, 136,
 137, 140, 147, 149, 152,
 161, 171, 174, 175, 186–
 190, 192, 193, 195, 197–
 199, 202–204, 216, 218,
 222, 228, 230, 235, 241,
 242, 276
 meaningless, 197
 morphographic theory, 186, 192
 morphological constituency, 202
 morphophonetic theory, 186, 198
 motivation
 phonetic, 215
 semantic, 215
 movable type, 333
muimi keitaiso, 197
 multigraph, 282
 musical notation, 5_{de}
- N
- n-component, 218, 219
 Nabatean, 245, 259
 NASA, 93
 Nashī, 260, 262
 nerve net, 82
 non-sexist writing, 73
 notation, 96
 nucleus, 148
 numeral system, 2_{de}
 NVME, 381
- O
- Obama, Michelle, 155
on, 191, 200
on'yaku, 191, 193
 1-dimensional graphetic
 sequence, 132
 onset, 29, 30, 148
 OPD, 277
 open syllable lengthening, 33
 OpenType, 22, 342, 343
 oracle bone, 229, 237
 orthographic
 change, 198
 depth, 297
 diglossia, 115
 evolution, 335
 influence, 288
 mistake, 118
 phenomenon, 62

- pluricentricity, 121
 - reform, 19, 329
 - regularization, 198
 - scheme, 343
 - strategy, 287, 318
 - style, 329, 333, 339, 345
 - tone bearing unit, 310, 314, 316
 - transfer, 283
 - variation, 197, 202, 319
 - word, 296, 304
 - orthography, 81, 92, 93, 97–99, 115, 121, 167, 188, 192, 247, 257, 270–273, 275–277, 283–291, 294–307, 311–320, 324, 325, 327, 333, 335, 337–339, 345, 346, 376
 - acceptability, 291
 - deep, 131
 - design, 273, 283, 290, 291
 - learnability, 290
 - reform, 301, 302, 305
 - reformed, 335, 338, 339, 342, 344–346
 - shallow, 131
 - strategy, 277, 281
 - tone, 296
 - traditional, 336, 338, 339, 342–346, 348–351
 - transparency, 290
 - unified, 286
 - Osmanlıca, 13_{de}
 - Oulipian constraint, 69
 - over-representation, 297
 - overdifferentiation, 277, 284, 286, 288
 - OWL, 131
- P
- p-frame, 369
 - Pamebbame* newspaper, 304
 - Papua New Guinea, 269–291
 - password typing, 354
 - Phoenician, 260
 - phone, 129, 367
 - phoneme, 11_{de}, 17, 44, 99, 102, 128, 129, 145, 147, 148, 171, 185, 188, 189, 192, 247, 260, 263, 269, 271, 275–278, 281–283, 285, 288, 290, 297–299, 303, 312–314, 317, 368
 - phonemic principle, 296
 - phonetic borrowing, 193, 202
 - phoneticity, 146
 - phonetics, 129
 - phonogram, 145, 210, 212, 218, 219
 - phonograph, 187, 192
 - phonographic principle, 260
 - phonology, 95
 - prosodic, 29
 - phylogeny, 376
 - pictogram, 146, 155, 158, 210, 212, 219, 222, 239
 - pistol emoji, 161
 - plain text, 138
 - polymorphemic word, 192
 - POP DIRECTIONAL FORMATTING, 150
 - Portuguese, 44–46, 50, 69, 80, 81, 84, 86, 191, 370, 372, 374, 381
 - POS tag, 137
 - POSIX, 82
 - presentation
 - form, 152
 - sequence, 156
 - Project Gutenberg, 371
 - prosodic hierarchy, 355, 357, 363
 - Proto-Sinaitic, 246
 - prototype, 105, 224
 - psycholinguistic reality, 273
 - punctuation, 18, 21, 22, 24, 82, 97, 106, 139, 143, 146, 147, 150, 151, 169, 171, 293–297, 299–305, 317–320, 371
 - pyromantic divination, 229

M component, 218

p-component, 218

Q

QID Emoji Proposal, 159

Qurʾan, 132, 133, 245, 259, 263

QWERTY, 14_{de}, 353

R

R, 357, 360

Rachana Aksharavedi, 342

RDF, 131

rebus, 191

redundancy, 9_{de}

regular

expression, 82–86

grammar, 148

planar, 148

language, 82

relation, 148

reliability island, 361, 363

rendering engine, 137, 149, 343

repertoire, 20, 98

replacement grapheme, 48

representative glyph, 144

revision strategy, 361

rich text, 131

RIGHT-TO-LEFT EMBEDDING, 150

Romaji-Kwai, 12_{de}

Roman

numeral system, 3_{de}

Rotokas, 275

ruby, 132

Russian, 19, 69, 171, 368, 370, 372,

374, 379, 381

S

s/p-component, 218

Saanich (SENĆOŦEN), 102

Saint-Marc Café, 149

Saussure, Ferdinand de, 10_{de}, 141,

175, 187, 363

schwa, 34, 331

script, 24, 91–99, 102, 104–107,

113–115, 117–123, 141–

144, 150, 151, 153, 162,

167, 171, 178, 180, 187, 189,

220, 229–235, 237, 238,

240, 241, 246–248, 258–

260, 263

Script Encoding Initiative, 180

scriptal pluricentricity, 119

script

alphabetic, 143

Arabic, 20, 106, 116, 117, 152,

154, 245, 247, 248, 259,

260, 263

Arkhanes, 120

broken, 103, 104, 138

cenemic, 118, 258

Cherokee, 98

Chinese, 19, 92, 106, 143, 171,

227, 228, 230, 232, 241,

242

Cretan, 120

Cyrillic, 91, 92, 105, 141

demotic, 119

Greek, 24, 93

hangul, 19

Hebrew, 24

hieratic, 118, 119

hieroglyphic, 119

hieroglyphic, 120

hiragana, 186

ideographic, 228

Japanese, 104

kanji, 185, 189

katakana, 186

Khmer, 147

Lao, 150

Latin, 20, 21, 91, 93, 97–

99, 102–107, 117, 141, 154,

186, 189, 289

Li, 232–234, 240

macro, 106, 107

Malayalam, 329–347

Nabatean, 245

- phonetic, 118, 228, 367
 pleremic, 118, 258
 Proto-Sinaitic, 246
 R, 360
 Roman, 92, 93, 106, 171, 178
 rōmaji, 186
 Seal, 231–235, 237–241
 Large, 230, 231
 Small, 231, 233
 semantic, 118
 semanto-phonetic, 228
 Semitic, 258
 sister, 106
 South-East Asian, 143
 Syriac, 143, 154, 245, 259
 Thai, 106, 150
 Ugaritic, 246, 253
 second articulation, 129, 140, 145
 segment, 305
 segmentation, 258
 semantic
 annotation, 131, 161
 component, 216, 218
 semanticity, 146
 semantic
 network, 218
 transparency, 195
 semasiography, 174
 semiographic principle, 258, 260,
 266
 semitic root, 260
 semivowel, 373
 sentence, 140
 separator grapheme, 49
 Serbian, 171, 371, 372, 381
 Serbo-Croatian, 117
 Service, 371
 set of signs, 19
 shoulder bone, 229
Shuōwén-jǐezì, 210, 218–220, 241
 sick
 i, 303
 v, 303
 signary, 188
 SIL International, 106, 269–272,
 276, 277, 281–283, 285–
 287, 290, 291, 294, 320
 SINGLE (method), 45
 Slovak, 142
 SMS, 176
 SOFT HYPHEN, 132, 142
 South Arabian, 247, 259
 Spanish, 44, 45, 47, 48, 50, 53, 66,
 69, 83, 86, 142, 370, 372,
 374, 381
 spectral decomposition, 367–386
 Spivak pronouns, 44, 51
 spyware, 355
 static underscore, 81
 stroke, 214, 233
 curved, 264
 leaning, 264
 order, 230
 thrown down, 264
 stylometry, 356
 Sudest, 275
 suffix order, 47, 66
 Sumerian, 114
 surface form, 148
 SVG, 133, 138
 Swedish, 370, 372, 378, 381
 syllabary, 385
 syllable, 20, 28–38, 77, 80, 132,
 146, 148, 150, 171, 192, 217,
 221, 228, 241, 242, 283,
 289, 299, 303, 311, 315,
 316, 353, 355, 357, 359
 peak, 29, 34
 synchronous computer-mediated
 communication, 354
 Syriac, 143, 150, 154, 155, 245, 259,
 263, 264
 M component, 218
 s-component, 218
- T
- tag character, 158, 160
 tag sequence, 158

Tagalog, 372, 381
 textual data, 139
 Thai, 106, 150
 time stamp, 354
 Times, 138
 tittle, 99
 Tok Pisin, 275, 283, 288
 token, 228
 tokples, 277, 287
 tone, 293, 305, 318
 marking, 294
 exhaustive, 294
 full, 294
 orthography, 296
 topogram, 145, 146
 tortoise shell, 229
 Trans New Guinea, 275, 278
 Turkish, 99, 102, 142, 152
 Twitter, 181
 type, 228
 typewriter, 130, 150, 152, 158, 286,
 294, 303, 334, 346
 Arabic, 151
 Malayalam, 336
 typography, 130, 132, 329, 339,
 342, 343

U

Ugaritic, 246
 umlaut, 293
 UN Declaration of Human Rights,
 274
 underdifferentiation, 277, 284,
 286, 288
 underlying form, 148
 underscore
 randomization, 81
 wandering, 82
 UNESCO, 273
 Unicode, 20–24, 66, 168, 175–181,
 233, 286, 304, 329, 336,
 339, 342–345
 aficionado, 162
 codespace, 141

 Consortium, 168, 177–181
 Uniskript, 272
 untextable letter, 284
 Urdu, 20, 371, 381
 user authentication, 354

V

v malade, 303
 variation
 diachronic, 198
 orthographic, 197
 Vietnamese, 10_{de}, 371, 372, 375,
 381
 virama, 332, 333, 336–339
 vowel, 29–38, 48, 50, 52, 53, 56,
 57, 72, 84, 102, 147, 150,
 188, 259, 263, 272, 281–
 283, 287, 288, 293, 294,
 296–303, 305, 309, 311,
 312, 317–319, 329, 331–
 333, 335–339, 357, 367,
 368, 373, 374, 376, 383
 binary, 30, 36, 37
 quantity, 27
 reduction, 31, 33
 shortening, 32
 unary, 30, 36
 Voynich manuscript, 369, 380

W

wandering underscore, 82
 WhatsApp, 168, 174, 176, 181
 Wikidata, 137, 159, 160
 Wikipedia, 98, 140, 170, 269, 371
 Wiktionary, 51, 55, 60, 67, 70, 71,
 76, 137
 WOMAN (emoji), 157
 word boundary, 264
 word
 avocalic, 380
 WordNet, 137, 161, 217
 writing system, 17–22, 24, 27, 28,
 37, 66, 82, 91–99, 102,

- 104–107, 112, 113, 115–118, 121–123, 127, 141, 151, 153, 154, 162, 167, 170, 171, 185, 187–189, 192, 193, 198, 199, 203, 209, 211, 224, 231, 246, 257–259, 266, 272, 286, 289, 291, 295, 329, 359, 367, 368, 370, 371, 374–376, 380, 382, 383, 385
- alphabetic, 93, 143, 368, 370, 371, 376
- Arabic, 258, 260
- Chinese, 189, 199, 211, 212, 220, 221, 224, 228, 241
- closed, 19–21, 24, 25
- Japanese, 91, 92, 185–187, 189, 192, 193, 201, 204
- morphemic, 188
- morphosyllabic, 228
- open, 19–22
- phonemic, 188
- semi-closed, 21
- semi-open, 20
- undeciphered, 367
- X**
- XHTML, 132, 133
- xíngsbū*, 232
- XML, 131, 138
- XSL-FO, 133
- Xu Shen, 210, 212, 213, 217, 219, 220, 227, 228, 241
- Y**
- Yago, 137
- yomi*, 190
- Z**
- Zen Buddhism, 196
- zero, 3_{de}
- ZERO WIDTH JOINER, 154, 157
- ZERO WIDTH NON-JOINER, 153–155
- ZWJ, *see* ZERO WIDTH JOINER