BRILL

ARABICA 69 (2022) 279-317

Arabica

brill.com/arab

Variation in the Use of Diacritics in Modern Typeset Standard Arabic: A Theoretical and Descriptive Framework

Andreas Hallberg[*] | ORCID: 0000-0001-9442-1495
Department of Languages & Literatures, University of Gothenburg,
Gothenburg, Sweden
andreas.hallberg@sprak.gu.se

Abstract

The extent to which the diacritic layer (*taškīl*) of the Arabic writing system is employed in modern typeset text differs considerably between genres and individual texts, with many in-between forms not aptly captured by the traditional binary categories of "vowelled" and "unvowelled" text. This article is the first to present a theoretical account of this variation applicable to modern typeset Standard Arabic. It is suggested that diacritics serve three basic functions: facilitation of reading comprehension; facilitation of prescriptively correct diction; and to evoke associations with other texts. Six modes of diacritization in modern typeset text are identified and related to data on rates of diacritization from a corpus of electronically published books. Further lines of research based on this framework are suggested.

Keywords

Modern Standard Arabic, orthography, *taškīl*, diacritics, corpus linguistics

Résumé

La façon dans laquelle la vocalisation (*taškīl*) du système d'écriture arabe est utilisée dans les textes modernes diffère sensiblement entre les genres et les textes, avec de

[*] I am grateful to Professor Tetz Rooke for pointing me to example (4) and to Lars-Dominik Braun for assistance in collecting the corpus material.

nombreuses formes intermédiaires qui ne sont pas adéquatement décrites par les catégories traditionnelles binaires de texte « vocalisé » et « non vocalisé ». Cet article est le premier à présenter un cadre théorique de cette variation qui soit applicable à l'arabe standard moderne. Il suggère que les signes diacritiques (au sens large, incluant aussi le *taškīl*) remplissent trois fonctions de base : faciliter la compréhension de la lecture, faciliter une diction formellement correcte et évoquer des associations avec d'autres textes. Six modes de diacritisation dans les textes modernes sont identifiés et mis en relation avec des données sur les taux de diacritisation provenant d'un corpus de livres publiés électroniquement. Il introduit enfin d'autres pistes de recherche s'appuyant sur ce cadre.

Mots clefs

Arabe standard moderne, orthographe, *taškīl*, diacritique, linguistique de corpus

Introduction

The Arabic language is written with an *abǧad*-type writing system[1] in which various phonological features, most importantly short vowels, are not indicated. Readers of Arabic have to provide this information themselves in order to produce complete phonological word forms from a written text. Similar to other *abǧad*s, the Arabic writing system has a subsidiary system of optional diacritics (*taškīl*) which may be used to indicate missing phonological information. A given word can thus be written in several ways depending on which diacritics are provided. The word *yudarrisu* 'he teaches,' for example, contains four short vowels and a lengthened consonant (*rr*), each of which may optionally be indicated by a diacritic. Some possible ways of writing this word are listed in (1),[*] all of which, except for (1a), are attested in *arabiCorpus*.[2] This variability is a striking feature of the Arabic writing system in that it deviates from

[1] Peter T. Daniels, "Fundamentals of Grammatology," *Journal of the American Oriental Society*, 110/4 (1990), p. 730.

[*] In the Arabic font used here, *kasra* is placed under the letter even when accompanied by *šadda*. In the quoted texts, *kasra* is in such cases placed under *šadda* above the letter.

[2] Dilworth B. Parkinson, "Under the Hood of arabiCorpus," in *Arabic Corpus Linguistics*, eds Tony McEnery, Andrew Hardie and Nagawa Younis, Edinburgh, Edinburgh University Press, 2019, p. 17-29.

the principle of *lexical consistency* typical of phonographic writing systems, that is, the principle that a given word is always written the same way.[3]

(1)
a. يُدَرِّسُ (all potential diacritics)
b. يُدرِّس
c. يُدَرس
d. يُدرس
e. يُدرّس
f. يدرّس
g. يدرس (no diacritics)

In scholarly literature, the Arabic writing system is typically described in largely binary terms as being either diacritized or undiacritized. This is usually referred to in terms of "vowelled" and "unvowelled" text, terms further discussed below. Typically, these descriptions state that diacritics are used in children's literature, poetry, and religious texts, sometimes with a mention in passing that occasional diacritics are sometimes also used in otherwise undiacritized text.[4] This binary description glosses over the many in-between forms of diacritization commonly used in Arabic texts, illustrated in (1) and in several examples below.

Stepping away from the binary paradigm to instead view diacritization as a complex, highly variant system raises a number of questions that have yet to be adequately addressed. How can we characterize different forms of diacritization in modern typeset Standard Arabic, whether on the word or text level? How do these forms relate to genre? Which diacritics are prioritized and for what purpose? How do the different forms of diacritization affect the reading

[3] Brett Kessler and Rebecca Treiman, "Writing Systems: Their Properties and Implications for Reading," in *The Oxford Handbook of Reading*, eds Alexander Pollatsek and Rebecca Treiman, New York, Oxford University Press ("Oxford Library of Psychology"), 2015, p. 13.

[4] See for example Eckehard Schulz, Günther Krahl and Wolfgang Reuschel, *Standard Arabic: An Elementary/Intermediate Course*, Cambridge, Cambridge University Press, 2000, p. 3; Karin C. Ryding, *A Reference Grammar of Modern Standard Arabic*, Cambridge-New York-Melbourne, Cambridge University Press, 2005, p. 30; and Elinor Saiegh-Haddad and Roni Henkin-Roitfarb, "The Structure of Arabic Language and Orthography," in *Handbook of Arabic Literacy: Insights and Perspectives*, eds Elinor Saiegh-Haddad and R. Malatesha Joshi, New York, Springer ("Literacy Studies," 9), 2014, p. 18.

process?[5] In order to investigate these questions we need a framework for understanding and categorizing diacritics in their various functions on the one hand, and on the other, the different in-between ways in which diacritics are employed in real-life texts. The aim of this article is to provide such a framework applicable to modern typeset Standard Arabic text. To the best of my knowledge, this is the first study on variation in the use of diacritics in modern typeset Arabic.

Patterns and variation in diacritization can be described on three levels of analysis: the level of individual diacritics, the word level, and the text level, with the conception of one level crucially depending on that of the former. This article roughly follows this order of levels of analysis. It is organized as follows. In the first section, definitions and terminological issues regarding typeset Arabic diacritics as a set of graphemes within the Arabic writing system are discussed. The second section discusses and delimits the inventory of diacritics in modern typeset text. The third section presents the scheme for categorizing diacritics in terms of their role in specifying and altering the phonological interpretation of written words. It is suggested that any given diacritic is either phonologically additive, specifying, or superfluous in relation to the letter sequence. In the fourth section, the main functions of diacritization as they relate to the reading process are discussed. Three such functions are proposed: facilitation of reading comprehension, facilitation of prescriptively correct diction, and an associative function. The fifth section presents a classificatory system for modern typeset text according to patterns of diacritization. This classification is illustrated and discussed in relation to data on rates of diacritization from a corpus of electronically published books. The article concludes with a summary and suggestions for future research based on this framework.

1 Definitions and Terminological Issues

A useful definition of a grapheme has recently been presented by Dimitrios Meletis.[6] According to this definition, a grapheme is a unit in writing that

[5] Experimental research on reading in Arabic has only recently begun to differentiate between diacritics that indicate case and mood inflection and those that are internal to the lexeme, typically by excluding the former in stimuli. See for example Elinor Saiegh-Haddad and Rachel Schiff, “The Impact of Diglossia on Voweled and Unvoweled Word Reading in Arabic: A Developmental Study from Childhood to Adolescence,” *Scientific Studies of Reading*, 20/4 (2016), p. 311-324. This, however, is only one of several fundamental distinctions of the functions of diacritics in relation to reading, as detailed below.

[6] Dimitrios Meletis, “The Grapheme as a Universal Basic Unit of Writing,” *Writing Systems Research*, 11/1 (2019), p. 26-49.

complies with the following three criteria: a) *lexical distinctiveness*, as tested with minimal pairs; b) it has a *linguistic value*, typically phonological; and c) *minimality*, it cannot be decomposed into other graphemes. By this definition, the modern Arabic script can be described as composed of three sets of graphemes: letters, diacritics, and punctuation (including word space). Letters are here defined as non-optional graphemes with an associated phonological value. This includes the graphemes enumerated in the alphabetic order, in addition to three letters not traditionally included in the alphabet (*tāʾ marbūṭa* ة, *alif maqṣūra* ى, and *hamza* ء, with its variants). Diacritics are here defined as optional, bound graphemes. They are bound in the sense proposed by Henry Rogers,[7] that is, they cannot be used independently and are always written in connection with a non-bound grapheme (a letter). Optionality and boundedness are thus the distinguishing features between letters and diacritics in modern typeset Arabic.

The differentiation between graphemes and phonemes is fundamental to modern linguistics. Yet it is often not adhered to in scholarly discussions on the Arabic writing system, leading to the establishment of some unfortunate terminology. The most common terms in the English literature for describing Arabic text with and without diacritics is "(un)vowelled" text, or alternatively, "(un)vowelized" text. This terminology is firmly established, but it is problematic. First, as a derivation of the word "vowel," it conflates the representation (a graphical entity) with what it represents (a phonetic entity). Second, some diacritics are either unrelated to vowels (e.g. *šadda* ـّ) or represent a vowel together with a consonant (*tanwīn* ـٌ, ـٍ, and ـً). Referring to the bound graphemes as "diacritics," and to texts as "diacritized" or "undiacritized," avoids these problems. This has the advantage of transparently referring to an orthographic property, unlike "vowels" or "vowelled."

The term "diacritic" is often used in Arabic linguistics, particularly in historical linguistics, for the dots that differentiate letters, for example the letters ج, ح, and خ, or ر and ز. The use of these dots has varied historically and they have been subject to omission, especially in handwritten text. For modern typeset text, however, they do not comply with the definition of diacritic given above: these dots are bound, but they are not optional. Furthermore, the letter-dots do not comply with Meletis' definition of a grapheme, in that they do not fulfil the criterion of having a linguistic value. There is no independent linguistic value for the dot in the letters ز, ن, and خ, for example. The use of the term "diacritic" for the letter-dots in contemporary typeset text is therefore deemed inappropriate by many scholars. Daniels, for example, argues that "since the

[7] Henry Rogers, *Writing Systems: A Linguistic Approach*, Malden, Blackwell Publishers ("Blackwell Textbooks in Linguistics," 18), 2005, p. 11.

dots are (now) integral parts of the letters, as much so as the dot on ⟨i⟩, they should not be given that label [‘diacritic’] in synchronic description.”[8] This restricted use of the term “diacritic” to the exclusion of letter-dots is also firmly established in Arabic natural language processing and corpus linguistics[9] and aligns with the Arabic term *taškīl* (the letter-dots are referred to in Arabic as *tanqīṭ* or *iʿǧām*).[10] It is adopted for the purposes of this article.

Before turning to the inventory of diacritics, there are two orthographic features that will not be dealt with in the following sections but that merit some discussion in that they, contrary to prescriptive orthography, are variably omitted in some contexts, giving them an ambiguous, in-between status between letters and diacritics. The first is the undotted *yāʾ* of the Egyptian orthographic tradition. In text published in Egypt “and in various regions where Egyptian spelling predominates,”[11] the letter *yāʾ* is in word-final position typically written without dots, giving the letter-forms ى and ـى which are identical to the letter *alif maqṣūra* (see [9] and [10] below for examples). This is a special feature of Egyptian orthography; when books originally published in Egypt appear in new editions outside of Egypt, the dots on word-final *yāʾ* are typically added for the text to comply with non-Egyptian orthographic norms.[12] Prescriptive rules on the matter are in Egyptian sources unclear or contradictory. In the government-issued Egyptian Arabic textbook for the first grade, the four forms of *yāʾ* are given as ى يـ ـيـ ـى,[13] as opposed to ي يـ ـيـ ـي elsewhere in the Arab world, and the dotted form of word-final *yāʾ* does not appear anywhere in the book, while in a spelling guide published in association with the Egyptian Academy the

[8] Peter T. Daniels, “The Arabic Writing System,” in *The Oxford Handbook of Arabic Linguistics*, ed. Jonathan Owens, Oxford-New York, Oxford University Press (“Oxford Handbooks in Linguistics”), 2013, p. 415.

[9] See, for example, Nizar Y. Habash, *Introduction to Arabic Natural Language Processing*, San Rafael, Morgan & Claywood (“Synthesis Lectures on Human Language Technologies,” 10), 2010, p. 7 for natural language processing; and Tressy Arts, Yonatan Belinkov, Nizar Habash, Adam Kilgarriff and Vit Suchomel, “arTenTen: Arabic Corpus and Word Sketches,” *Journal of King Saud University – Computer and Information Sciences*, 26/4 (2014), p. 358, n. 7 for corpus linguistics.

[10] Saiegh-Haddad and Henkin-Roitfarb, “The structure of Arabic language and orthography.”

[11] Timothy Buckwalter, “Issues in Arabic Morphological Analysis,” in *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, eds Abdelhadi Soudi, Antal van den Bosch and Günter Neumann, Dordrecht, Springer (“Text, Speech and Language Technology,” 38), 2007, p. 30.

[12] Compare, for example, Yūsuf Idrīs, *al-Ab al-ġāʾib*, Cairo, Maktabat Miṣr, 1987 and Yūsuf Idrīs, *al-Ab al-ġāʾib*, Windsor, Hindawi, 2017.

[13] Muḥammad Ṣalāḥ Farağ and Muḥammad ʿAbd al-Ḥamīd Ġurāb, *Hayyā natawāṣal: al-luġa l-ʿarabiyya li-l-ṣaff al-awwal al-ibtidāʾī, al-faṣl al-dirāsī l-ṯānī*, Cairo, Wizārat al-tarbiya wa-l-taʿlīm, 2008, p. 89.

dotted form in word-final position is prescribed.[14] In some texts published in Egypt, both the dotted and the un-dotted forms are used, seemingly at random. For example, in the best-selling novel *Šīkāğū* by the Egyptian author ʿAlāʾ al-Aswānī[15] the preposition *fī* 'in' (*n* = 2,138) and the pronoun *hiya* 'she' (*n* = 169) are spelled with the dotted *yāʾ* 3% and 15% of the time respectively.[16] The reasons for this variation are unclear. The diacritics listed in following section (*ḥarakāt*, *šadda*, etc.) are in the novel used primarily to disambiguate homographs or opaque orthographic forms, as shown below. One is hard-pressed to find a similar reason behind the occasional addition of the dots on word-final *yāʾ*. Nevertheless, this variable addition of the letter-dots for *yāʾ* makes for an in-between case of letter-dots that show diacritic-like optionality, if only in the Egyptian orthographic tradition.

The second feature with an ambiguous status as letter or diacritic is stem-initial *hamza*. Prescriptively, stem-initial *hamza* is either a) *hamzat al-qaṭʿ*, realized as /ʔ/ and is written أ or إ; or b) *hamzat al-waṣl*, which has no phonetic realization when preceded by a vowel and written with bare *alif* ا or optionally with *waṣla* ٱ. Orthographic rules leave little doubt as to where either form applies.[17] From a prescriptive-normative standpoint, then, regular *hamza* أ/إ indicating *hamzat al-qaṭʿ* is not optional and is therefore not a diacritic, while *waṣla* ٱ marking *hamzat al-waṣl* is optional and therefore a diacritic. However, contrary to prescriptive rules, regular *hamza* is sometimes omitted in typeset text, in particular in print news.[18] For example, in the Newspaper section of arabiCorpus, representing print news from 1996-2012,[19] the word *ağl* أجل 'sake' (as in 'for the sake of'), without clitics, is written without the *hamza* (اجل)

[14] Ayman Amīn ʿAbd al-Ġanī, *al-Kāfī fī qawāʿid al-imlāʾ wa-l-kitāba*, Cairo, Dār al-tawfīqiyya li-l-turāṯ, 2012, p. 31, 103.

[15] ʿAlāʾ al-Aswānī, *Šīkāgū*, Cairo, Dār al-šurūq, 2007.

[16] Data extracted from arabiCorpus.

[17] See Ryding, *A Reference Grammar of Modern Standard Arabic*, p. 16-21 for an overview, and ʿAbd Allāh b. ʿAqīl, *Šarḥ Ibn ʿAqīl ʿalā Alfiyyat Ibn Mālik*, ed. Anṭūniyūs Buḍrus, Tripoli, al-Muʾassasa l-ḥadīṯa li-l-kitāb, 2005, p. 530-531; and Anṭwān al-Daḥdāḥ, *Muʿğam qawāʿid al-luġa l-ʿarabiyya fī ğadāwil wa-lawḥāt*, Beirut, Maktabat Lubnān, 1992, p. 12-13, for examples of classical and modern descriptions of these rules in the Arabic grammatical tradition.

[18] According to El-Said M. Badawi, Michael G. Carter and Adrian Gully, *Modern Written Arabic: A Comprehensive Grammar*, London-New York, Routledge ("Routledge Comprehensive Grammars"), 2004, p. 12, the opposite is also common, i.e. the insertion of regular *hamza* where *hamzat al-waṣl* is prescribed. This is, however, not borne out in inquiries. Of all instances in arabiCorpus of the three words given in *Modern Written Arabic* as examples of this phenomenon, only a small proportion are written with addition of prescriptively incorrect regular *hamza*: 0.6% of الانتظار 'the waiting' (*n* = 9,835), 0.6% of انهيار 'collapse' (*n* = 11,811), and 0.4% of اتباع 'following' (*n* = 11,612).

[19] Parkinson, "Under the hood of arabiCorpus," p. 19.

34% of the time (n = 65,880). In other types of text, the corresponding percentage is significantly lower. In the Modern Literature section of the same corpus the corresponding percentage is 2% (n = 453), and in the Arabic News Text Corpus,[20] a corpus of electronically published news articles from 2021, it is 0.6% (n = 4,717). Similarly, in the examples below, all from book-length texts, there is no example of *hamza*-omission. These data indicate that frequent *hamza*-omission is restricted to print news and that news text to a large extent has been standardized to conform with prescriptive norms in the transition to digital text production.[21]

The dots of the word-final *yāʾ* and the stem-initial *hamza* are thus edge-cases of diacritics: they are subject to optional omission, as are diacritics, but only in specific contexts, in texts published in Egypt and in print news respectively, and their omission is prescriptively incorrect (*hamza*-omission) or of unclear prescriptive status (undotted *yāʾ*). As such, they contrast with the "core" diacritics listed below which are optional in typeset text throughout the Arabic speaking world and whose omission is prescriptively sanctioned. These core diacritics are the focus of the rest of this article.

2 The Diacritic Inventory

Below is an annotated list of the inventory of diacritics used in modern typeset Arabic (excepting the two ambiguous edge-cases discussed above). The first three (*ḥarakāt*, *sukūn*, and *alif ḫanǧariyya*) indicate vowel phonemes or lack thereof, while the others serve diverse functions defying any straightforward

[20] Amina Chouigui, Oussama Ben Khiroun and Bilel Elayeb, "An Arabic Multi-Source News Corpus: Experimenting on Single-Document Extractive Summarization," *Arabian Journal for Science and Engineering*, 46/4 (2021), p. 3925-3938.

[21] The practice of *hamza*-omission in print news likely originates from the hot-metal typecasting technology that dominated news-paper production for much of the 20th century. Restrictions of this technology required the large number of sorts used for traditional manual Arabic typesetting to be drastically reduced, and early Arabic fonts for hot-metal typecasting did not feature dedicated characters for أ or إ, as described by Titus Nemeth, "Arabic Hot Metal," *Philological Encounters*, 3/4 (2018), p. 496-523. Even when these characters became available, they were often not used, as can be seen in several examples in Titus Nemeth, "Simplified Arabic: A New Form of Arabic Type for Hot-Metal Composition," in *Typography Papers 9*, eds Eric Kindel and Paul Luna, London, Hyphen, 2013, p. 173-189, and the practice seems to have survived even after hot-metal typesetting became obsolete, as shown above. Nemeth, "Simplified Arabic," p. 176, notes that traditional manual typesetting remained the norm in book publishing. This may explain the difference in *hamza*-omission between print news and literature shown above, in that there was no technological restriction in the production of literary book-length texts instigating the practice of *hamza*-omission.

categorization, indicating a subset of the case-inflecting morphemes (*tanwīn*), consonant lengthening (*šadda*), and a specific prosodic-phonetic feature (*waṣla*). The diacritics are listed with their Arabic terms. Note that all these terms are intended to refer to the diacritic, and not to the phonological phenomenon it represents. For example, *šadda* is a diacritic representing the phonological phenomenon of consonant lengthening. The Arabic term is followed by the English term where available. *Sukūn*, *šadda*, and *waṣla* lack English terms and the Arabic terms are generally used also in English contexts, a practice followed here. Besides the diacritics listed here, there is an additional set of diacritics specific to the Qurʾān, relating to specific features of recitation, such as pauses and assimilation.[22] Again in the Qurʾān, some diacritics are used in ways that differ from other texts, such as the use of *sukūn* to signal that a final letter *alif* is silent. These will not be discussed further here since they are not part of the modern orthography.

- *Ḥarakāt* (sg. *ḥaraka*), vowel diacritics ـَ (*fatḥa*), ـُ (*ḍamma*), and ـِ (*kasra*) provide only vowel information. They do this in one of two ways. In most instances they represent a short vowel (/a/, /i/ or /u/). When followed by the letter that represents the respective long vowel (ـَا /aː/, ـُو /uː/, and ـِي /iː/), they serve to disambiguate the phonetic value of that letter as representing a long vowel rather than a consonant. We can thus make a further distinction between *short* and *long* vowel diacritics in terms of their function.
- *Sukūn* ـْ indicates the absence of a vowel after a consonant.
- *Alif ḫanǧariyya*, dagger *alif* ـٰ indicates a long vowel /aː/. It is a remnant of Qurʾānic orthography and in modern Arabic only appears in a handful of words.[23] Except for the word *Allāh*, which is often represented by a multi-letter glyph that includes the dagger *alif* (ﷲ), it is, however, rarely used even in fully diacritized texts, and is instead substituted by ـَا.
- *Tanwīn*, nūnation diacritics ـً, ـٌ, and ـٍ represent the phonemic pair of a vowel and the consonant /n/. These diacritics only occur at the end of words, to indicate case inflecting morphemes. ـً and ـٍ take the form of the doubled associated vowel diacritic, but are generally conceptualized as being one single grapheme, notably in the Unicode standard.[24] They are therefore not decomposable in typewritten text and can be considered to have graphemic status.[25]

[22] Kristina Nelson, *The Art of Reciting the Qurʾan*, Cairo-New York, The American University in Cairo Press, 2001, p. 28.

[23] Ryding, *A Reference Grammar of Modern Standard Arabic*, p. 28.

[24] *The Unicode Standard: Version 12.0 – Core Specifications*, Mountain View, Unicode Consortium, 2019, p. 368.

[25] See Meletis, "The Grapheme as a Universal Basic Unit of Writing," p. 36.

- *Šadda* ـّ indicates the phonemic lengthening of a consonant. There is no diacritic to indicate the absence of this feature, in analogy with how *sukūn* indicates the absence of a vowel. However, if a vowel or nūnation diacritic is added to a letter, a potential *šadda* on that letter is obligatory. The presence of a vowel or nūnation diacritic without an accompanying *šadda* (e.g. درَس) is thus interpreted as indicating the absence of consonant elongation.
- *Waṣla* ٱ is only used in connection with the letter *alif* ا in word-initial position to indicate that the letter does not have a phonetic realization.

3 The Roles of Diacritics in Phonemic Specification

The primary function of Arabic diacritics is to phonemically disambiguate the letter sequences of a word. From this perspective, they can be categorized into three groups according to how they phonemically relate to the letter to which they are attached:

- *Phonemically additive diacritics* are those that do not alter or specify the phonemic value of a letter but add phonemic information independent of the letter. This group includes *sukūn*, the short vowel diacritics, and nūnation diacritics. The nūnation diacritic ـً is phonemically specifying (see below) when accompanying a word final ا, which ambiguously represents either the suffix /an/ or a long vowel /aː/.
- *Phonemically specifying diacritics* are those that disambiguate or change the default phonemic value of a letter. This group includes *šadda* ـّ, which specifies the phonemic value of the letter as lengthened; *waṣla* ٱ, which specifies a letter ا as phonemically void; and the long vowel diacritics ـُ and ـِ, which specify the phonemic values of the letters و and ي respectively as long vowels. (For the long vowel diacritic ـَ, see below.)
- *Phonemically superfluous diacritics,* lastly, are diacritics that do not add any phonemic information and do not disambiguate any letters. This includes the vowel diacritic ـَ preceding one of the letters ة, ى, or word internal ا. For word final ا, the preceding ـَ is not phonemically superfluous, but specifying, as it excludes the reading of this letter as representing the nūnation /an/.

4 Functions of Diacritization in Reading

Reading, the extraction of linguistic information from text, crucially depends on word identification, that is, the association by the reader of a letter string with an entry in the mental lexicon.[26] It is with this sub-process of reading that diacritics primarily interact. In this section, a theoretic account of word identification, the dual-rout model of reading, is presented, where three basic functions of reading are lined out and related to this model.

For efficient reading, the process of word identification must be largely automatic to be quick enough to allow for a large enough number of words to be stored in the short-term memory for processing. The short-term memory is primarily a phonological storage mechanism, and the written word must therefore be converted to a phonological code.[27] Phonological recoding is therefore central in silent reading as well, as has been amply demonstrated in a number of experimental paradigms.[28]

There are several theoretical models of visual word recognition and of how the phonological code is generated from the written input. The majority of specialists ascribe to some form of the so-called dual-route model.[29] This refers to the idea that there are two separate routes from the written word to the lexical entry in the mental lexicon and its associated phonological form: the *indirect route* and the *direct route*. In the indirect route (also known as the non-lexical route or assembled phonology), the phonological form of the word is assembled by applying letter-to-sound conversion rules to the written word. In the direct route (also known as the lexical route or address phonology), the written word is processed as a single non-decomposable sign that is matched to a complete phonological word. Both the direct route and the indirect route probably play some role in skilled reading. Without the indirect route, readers would not be able to construct phonological forms for words that are new to them or that they have not previously encountered in writing. Without the direct route, readers would not have any way to produce the correct phonological code for irregularly spelled words, such as the English word *have*, for which regular letter-to-sound conversion rules would produce an incorrect phonological form. The direct route is, furthermore, significantly quicker, since

[26] See Keith Rayner, Alexander Pollatsek, Jane Ashby and Charles Clifton, *Psychology of Reading*, New York-London, Prentice Hall, 2012, chap. 3 for an in-depth discussion.

[27] Alexander Pollatsek, “The Role of Sound in Silent Reading,” in *The Oxford Handbook of Reading*, eds Alexander Pollatsek and Rebecca Treiman, New York, Oxford University Press (“Oxford Library of Psychology”), 2015, p. 194.

[28] See Rayner, Pollatsek, Ashby and Clifton, *Psychology of Reading*, p. 137-155 and Pollatsek, “The role of sound in silent reading” for reviews.

[29] See Max Coltheart, “Dual Route and Connectionist Models of Reading: An Overview,” *London Review of Education*, 4/1 (2006), p. 5-17, for an introduction to the model; and Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon and Johannes C. Ziegler, “DRC: A Dual Route Cascaded Model of Visual Word Recognition and Reading Aloud,” *Psychological Review*, 108/1 (2001), p. 204-256, for an in-depth discussion.

it does not involve any analysis or application of rules, and for most words it wins out in more quickly outputting a phonological form. Developing the direct route is therefore crucial for effective reading. The typical development of reading skills involves a phase of heavy reliance on the less efficient indirect route, where the reader “sounds out” words, with more and more words becoming accessible to the direct route as reading skills develop.[30]

Since the Arabic writing system is phonemically impoverished, the indirect route is sometimes unreliable or unavailable. However, with the Arabic morphological system of roots and patterns,[31] a phonological form can often be produced even for an unfamiliar undiacritized word via the indirect route, as its form is matched with a known pattern. This somewhat complicates the interpretation of the dual-route model as it relates to Arabic.[32] For example, if a reader encounters the derived word ضالعين ‘crooked, involved (in a plot)’ that is unfamiliar to them, they can reliably match this to the known pattern $R_1\bar{a}R_2iR_3\bar{\imath}n$, the pattern of an active masculine plural participle in form I, and mesh this form with the phonological information provided by the written word to produce the phonological word *ḍāliʿīn*. This process requires extensive knowledge of available patterns. These become increasingly available as knowledge of this system develops, in what David L. Share and Amalia Bar-On[33] has called the *lexico-morpho-orthographic* phase of reading development in *abǧad*-type writing systems. Unfamiliar non-derived words are more problematic for the indirect route since the vowel pattern in these words are much less predictable. For example, an unfamiliar word ضلع ‘to be/become crooked,’ a non-derived word with the root from the example above, potentially fits some twenty odd patterns ($R_1aR_2aR_3$, $R_2aR_3R_4$, $R_1uR_2aR_3$, $R_1iR_2R_3$, $R_1uR_2uR_3$, $R_1aR_2iR_3$, etc.). For such words, diacritics may helpfully be added to indicate the intended pattern.

Thus, the addition of diacritics makes the indirect route available or more reliable in outputting the correct form, to the benefit of the novice reader,

[30] Margaret Harris and Max Coltheart, *Language Processing in Children and Adults: An Introduction*, London-New York, Routledge (“Introductions to Modern Psychology”), 1989, p. 85-99.

[31] Ryding, *A Reference Grammar of Modern Standard Arabic*, p. 35-39.

[32] For further discussion on this point, see Miriam Taouka and Max Coltheart, “The Cognitive Processes Involved in Learning to Read in Arabic,” *Reading and Writing*, 17/1 (2004), p. 27-57; and David L. Share, “On the Anglocentricities of Current Reading Research and Practice: The Perils of Overreliance on an ‘Outlier’ Orthography,” *Psychological Bulletin*, 134/4 (2008), p. 584-615.

[33] David L. Share and Amalia Bar-On, “Learning to Read a Semitic Abjad: The Triplex Model of Hebrew Reading Development,” *Journal of Learning Disabilities*, 51/5 (2018), p. 444-453.

who may not have developed the direct route for a specific word. For the proficient reader, however, diacritization often "summons undue phonological recoding"[34] as it presents the word in what is in effect a novel or unusual form, possibly blocking the direct route and forcing the reader to employ the less efficient phonological parsing of the indirect route. This is shown in the fact that skilled readers read text with extensive diacritization more slowly than they do undiacritized text, as has been demonstrated in numerous studies. For example, Haitham Taha[35] had 2nd, 4th, and 6th graders read word lists with and without diacritics. For the 2nd graders there was no difference in reading speed or accuracy between the two conditions, but the 4th and 6th graders read the word lists without diacritics both faster and with higher accuracy. The same result was found by Raphiq Ibrahim[36] and Aula Khateeb Abu-Leil, David L. Share, and Raphiq Ibrahim[37] for 8th graders. For reading of continuous texts, Ali Al Midhwah and Mohammad T. Alhawary[38] found similar results for university students of Arabic as a second language, as did G. Roman and B. Pavard[39] and Ehab W. Hermena *et al.*[40] for adult native-speaking readers. The latter two, furthermore, measured eye movements during reading, and showed that the increased reading times in reading diacritized text were related to more and longer fixations, a measure known to be related to word identification.[41] Accordingly, even though diacritics may facilitate word identification for some

[34] Saiegh-Haddad and Schiff, "The Impact of Diglossia on Voweled and Unvoweled Word Reading in Arabic."

[35] Haitham Taha, "Deep and Shallow in Arabic Orthography: New Evidence from Reading Performance of Elementary School Native Arab Readers," *Writing Systems Research*, 8/2 (2016), p. 133-142.

[36] Raphiq Ibrahim, "Reading in Arabic: New Evidence for the Role of Vowel Signs," *Creative Education*, 4/4 (2013), p. 248-253.

[37] Aula Khateeb Abu-Leil, David L. Share and Raphiq Ibrahim, "How Does Speed and Accuracy in Reading Relate to Reading Comprehension in Arabic?," *Psicologica: International Journal of Methodology and Experimental Psychology*, 35/2 (2014), p. 251-276.

[38] Ali Al Midhwah and Mohammad T. Alhawary, "Arabic Diacritics and Their Role in Facilitating Reading Speed, Accuracy, and Comprehension by English L2 Learners of Arabic," *The Modern Language Journal*, 104/2 (2020), p. 418-438.

[39] G. Roman and B. Pavard, "A Comparative Study: How We Read in Arabic and French," in *Eye Movements from Physiology to Cognition*, eds John Kevin O'Regan and Ariane Levy-Schoen, Amsterdam-New York, North-Holland, 1987, p. 431-440.

[40] Ehab W. Hermena, Denis Drieghe, Sam Hellmuth and Simon P. Liversedge, "Processing of Arabic Diacritical Marks: Phonological/Syntactic Disambiguation of Homographic Verbs and Visual Crowding Effects," *Journal of Experimental Psychology: Human Perception and Performance*, 41/2 (2015), p. 494-507.

[41] Rayner, Pollatsek, Ashby and Clifton, *Psychology of Reading*, p. 236.

words, heavy use of diacritics has, averaging across all words, an impeding effect in skilled reading.

With this background in mind, we can describe diacritization in modern typeset Standard Arabic as potentially serving three basic functions: to facilitate word identification, to facilitate correct diction, and as having an associative function unrelated to the reading process *per se*. These three functions are discussed in the remainder of this section.

4.1 *Facilitation of Word Identification*

In undiacritized Arabic text homographs abound. By one estimate, as many as every third word in normal text is homographic.[42] Arabic homographs are virtually always heterophonic and can therefore be disambiguated with diacritics that indicate the phonological form. Often, the addition of one well placed diacritic is enough to uniquely identify the intended reading of a homograph. Furthermore, homographs can be further divided into *lexical* homographs, where the letter string corresponds to two or more different lexemes (e.g. درس *dars* 'lesson' and *daras* 'studied [3rd m. sg.]'), and *inflectional homographs*, where the letter string corresponds to two or more inflectional forms of the same lexeme (e.g. درست 'studied' for 1st sg., 2nd m. sg., 2nd f. sg., and 3rd f. sg.). Even more common is homography in case and mood inflection. Case and mood inflection, however, plays little to no role in reading comprehension, and is therefore discussed below as relating to correct diction.

The second situation where diacritics may facilitate word identification is in inflectional forms where one of the letters representing a root consonant is omitted, giving an opaque orthographic form. In some situations, the opaque orthographic form is enough to merit diacritization, even if the word is not homographic. A common example is in the so-called defective nominals where the final root consonant و or ي is omitted in the indefinite masculine singular form making the word difficult to identify (e.g. نواد for *nawādin* 'clubs'). Such words are often supplied with a nūnation diacritic specifying the phonological form of the word (نوادٍ), thereby facilitating word identification.

The third situation where diacritics may facilitate reading for proficient readers is in rare Arabic words that may be unfamiliar to the reader, that is, words not represented in their internal lexicon and where the reader must therefore revert to the indirect route to produce a phonological representation. Examples of this were presented above.

[42] Salim Abu-Rabia, "Reading Arabic Texts: Effects of Text Type, Reader Type and Vowelization," *Reading and Writing*, 10/2 (1998), p. 105-119.

Loanwords and non-Arabic proper names are a special case in terms of how they are represented in the Arabic writing system in that they are diacritized by different principles than are native Arabic words. It might be expected that potentially unfamiliar non-Arabic words will often be provided with diacritics in analogy with rare Arabic words. This, however, is not the case. Instead, non-Arabic words are generally written based on different principles than are words of Arabic origin. In foreign words, all or most vowels are represented by letters, as if they were all long vowels.[43] In effect, non-Arabic words are written with the Arabic script employed as an alphabetic writing system, where all phonemes, including vowels, are represented by a letter, rather than as an *abǧad*, where many vowels are not. This is exemplified in (2). This principle is not consistently applied to all words, giving rise to variants in spelling, for example in (2c),[44] which may reflect regional differences.

(2)
a. فيلم *film* 'film'
b. فيلا *filla* 'villa'
c. كمبيوتر / كومبيوتر *kumbyūtar* 'computer'
d. بوش 'Bush'
e. أوباما 'Obama'
f. ترامب 'Trump'

The convention of using alphabetic principles rather than diacritics in the spelling of foreign words limits the phonemic accuracy by which these words can be represented. Since all vowels in the source language are effectively presented as long vowels, the phonological form provided in the written word is often distorted with respect to the source language. In many cases, this could have been avoided if diacritics were used. For example, the Arabic orthographic representation of the American proper name Bush (2d) leads to a phonetic rendering with a long vowel /buːʃ/, which is a common pronunciation of this name in Arabic, but which does not correspond to the pronunciation in the source language. Diacritics would have made the orthographic form بُش available, representing the phonological form /buʃ/, which is closer to the original English pronunciation. This option, however, is not available given the

[43] Badawi, Carter and Gully, *Modern Written Arabic*, p. 18.

[44] Tim Buckwalter and Dilworth B. Parkinson, *A Frequency Dictionary of Arabic: Core Vocabulary for Learners*, London-New York, Routledge ("Routledge Frequency Dictionaries"), 2011, p. 169.

convention of rendering foreign words with alphabetic rather than abǧadic orthographic principles.

4.2 *Facilitation of Correct Diction*

A second and distinctly different function of diacritics is to facilitate prescriptively correct diction. Arabic diglossia,[45] combined with the phonemically impoverished *abǧad* writing system, makes for ample variation in reading aloud. Many written words have standard and non-standard cognates,[46] and it is thus possible to read many words aloud in either a standard or a non-standard form. There are several aspects of the phonological forms of Arabic words that are specific to Standard Arabic and that may be specified with diacritics, but that are marginal or irrelevant for reading comprehension since they do not alter or specify the meaning of the word. As such, they are points of differentiation between, on the one hand, formal, prescriptively correct styles of reading aloud, and informal styles that are closer to the non-standard natively spoken Arabic variety of the reader on the other. This informal style has been described by a number of scholars[47] but has yet to be systematically investigated. Where the author or editor intends for a text to be read aloud with prescriptively correct formal diction, diacritics representing these standard features may be added to help the reader do so. These features include

- case and mood inflection (e.g. the endings in *al-dars-u/-a/-i* 'the lesson[-nom./-acc./-gen.]' or *yadrus-u/-a/-ø* 'he studies[-ind./-sub./-jus.]');
- final short vowels in function words (e.g. in *ʿinda* 'with' or *bayna* 'between');
- short word-internal vowels in words with commonly used non-standard vowel patterns (e.g. *minṭaqa* 'area' instead of the non-standard *manṭiqa*, or *miṣr* instead of the non-standard *maṣr* 'Egypt');

[45] Charles A. Ferguson, "Diglossia," *Word*, 15/2 (1959), p. 325-340; *id.*, "Epilogue: Diglossia Revisited," in *Understanding Arabic: Essays in Contemporary Arabic Linguistics in Honor of El-Said Badawi*, ed. Alaa Elgibali, Cairo, American University in Cairo Press, 1996, p. 49-67.

[46] See Elinor Saiegh-Haddad and Bernard Spolsky, "Acquiring Literacy in a Diglossic Context: Problems and Prospects," in *Handbook of Arabic Literacy: Insights and Perspectives*, eds Elinor Saiegh-Haddad and R. Malatesha Joshi, Dordrecht, Springer ("Literacy Studies," 9), 2014, p. 225-240, for estimates of the extent of this phenomenon.

[47] Mohamed Maamouri, *Language Education and Human Development: Arabic Diglossia and Its Impact on the Quality of Education in the Arab Region*, discussion paper prepared for The World Bank, The Mediterranean Development Forum, Marrakech, 3-6 September 1998, Philadelphia, International Literacy Institute, 1998, p. 43; Dilworth B. Parkinson, "Searching for Modern Fuṣḥa: Real-Life Formal Arabic," *al-ʿArabiyya*, 24 (1991), p. 38; Badawi, Carter and Gully, *Modern Written Arabic*, p. 33; Andreas Hallberg, *Case Endings in Spoken Standard Arabic: Statistics, Norms, and Diversity in Unscripted Formal Speech*, Doctoral dissertation, Lund University, 2016, p. 79.

- elongation of the consonant in the so-called *nisba* ending *-iyy* (e.g. *ʿarabiyy* instead of the non-standard *ʿarabi* 'Arabic').

Of these features, case and mood inflection has an especially important status in the language community as a marker of linguistic correctness. It is a focal point of the Arabic grammatical tradition[48] and of Arabic language instruction in the Arab world.[49] Case and mood inflection is syntactically redundant due to the fixed word order in Arabic,[50] and is not phonologically encoded in skilled silent reading, or only to a very limited extent.[51]

Diacritics indicating these standard features limit the available forms of phonological encoding of these words to the option traditionally considered correct. These diacritics may therefore impede word identification, since "the assembled phonological form of the words does not, in many cases, align with the SpA [Spoken Arabic] form that speakers harbor in their lexicons,"[52] allowing access to these words only through the L2 Standard Arabic lexicon. The function of these diacritics can therefore be seen as facilitating correct diction by attempting to actively block words from being phonologically recoded according to non-standard morphology.

[48] Kees Versteegh, "Arabic Grammar and Corruption of Speech," *al-Abḥāṯ*, 31 (1983), p. 139-160; Georges Bohas, Jean-Patrick Guillaume and Djamel Eddine Kouloughli, *The Arabic Linguistic Tradition*, London-New York, Routledge ("Arabic Thought and Culture"), 1990, p. 10, 50.

[49] Muhammad H. Ibrahim, "Linguistic Distance and Literacy in Arabic," *Journal of Pragmatics*, 7/5 (1983), p. 512; Maamouri, *Language Education and Human Development*, p. 53; Niloofar Haeri, *Sacred Language, Ordinary People: Dilemmas of Culture and Politics in Egypt*, New York-Basingstoke, Palgrave Macmillan, 2003, p. 40; Allon J. Uhlmann, "Arabs and Arabic Grammar Instruction in Israeli Universities: Alterity, Alienation and Dislocation," *Middle East Critique*, 21/1 (2012), p. 101-116.

[50] Clive Holes, *Modern Arabic: Structures, Functions, and Varieties*, Washington, Georgetown University Press ("Georgetown Classics in Arabic Language and Linguistics"), 2004, p. 17, 173.

[51] This has been the assumption among linguists, for example in Mary Catherine Bateson, *Arabic Language Handbook*, Washington, Center for Applied Linguistics ("Language Handbook Series"), 1967, p. 81-82; Jaroslav Stetkevych, *The Modern Arabic Literary Language: Lexical and Stylistic Developments*, Washington, Georgetown University Press ("Georgetown Classics in Arabic Language and Linguistics"), 2006, p. 84; and Elinor Saiegh-Haddad, "MAWRID: A Model of Arabic Word Reading in Development," *Journal of Learning Disabilities*, 51/5 (2018), p. 454-462. For recent experimental evidence supporting this, see Taouka and Coltheart, "The Cognitive Processes Involved in Learning to Read in Arabic"; and Andreas Hallberg and Diederick C. Niehorster, "Parsing Written Language with Non-Standard Grammar: An Eye-Tracking Study of Case Marking in Arabic," *Reading and Writing*, 34/1 (2021), p. 27-48.

[52] Saiegh-Haddad and Schiff, "The Impact of Diglossia on Voweled and Unvoweled Word Reading in Arabic," p. 9.

4.3 *Associative Uses of Diacritics*

Diacritics that neither facilitate reading comprehension nor prescriptively correct diction can generally be described as serving an *associative function*. Diacritics have an associative function when their purpose is to evoke associations with text types associated with diacritization and thereby affect the reader's attitude towards the text. Full diacritization is primarily associated with the Qurʾān, the first written text many novice readers of Arabic encounter, but also with classical poetry and hadith, as well as other texts of the literary and religious canon. All these texts are highly revered, regarded as representing the literary and linguistic ideal, and are often memorized in order to be recited.[53] Extensive diacritization thus carries two associations: on the one hand with cultural centrality and reverence, and on the other with the reading practices of memorization and recitation. Authors may capitalize on these associations by adding diacritization to a text in order to evoke one or both of these associations.

As a function that capitalizes on the associations of the visual characteristic of the text rather than on specific aspects of the reading process, this function is realized on the textual level, rather than on the diacritic or word level. That is, for the associative function to be achieved, enough diacritics need to be added for the text to be visually similar to culturally revered diacritized texts. This may explain the many seemingly random uses of diacritics often found in texts with a heavy use of diacritics, where, for example, a reoccurring word is diacritized differently on different occasions. Some examples of this are mentioned below. The function of individual diacritics for word identification and reading comprehension are of little importance for this function to be achieved, and authors may add diacritics haphazardly to reach a given threshold.

One example of the associative function of diacritics to evoke reverence or respect is the novel *al-Sīratān* by Salim Barakat (b. 1951). This novel, and others by the same author, are unusual in being heavily diacritized, as shown in (3), the first paragraph of this novel.[54] The novel employs the mode of *simplified diacritization* described below.

[53] Daniel A. Wagner and Abdelhamid Lotfi, "Learning to Read by 'Rote,'" *International Journal of the Sociology of Language*, 42 (1983), p. 111-121; Haeri, *Sacred Language, Ordinary People*; Kees Versteegh, "Learning Arabic in the Islamic World," in *The Foundations of Arabic Linguistics III: The Development of a Tradition: Continuity and Change*, eds Georgine Ayoub and Kees Versteegh, Leiden-Boston, Brill ("Studies in Semitic Languages and Linguistics," 94), 2018, p. 245-267.

[54] Translations are my own unless otherwise indicated.

> (3)
>
> ما ٱلَّذي تراه؟ قُلْ لي أيُّها الطِّفْلُ ما ٱلَّذي تراه؟ هَضَبَتانِ في ٱلأُفُقِ، وَعِقْدٌ مِنَ ٱلقُرى
> وَتُرابٌ يَتَرَنَّحُ بَيْنَ صَيْفٍ طائِشٍ وَبَيْنَ شِتاءٍ أَحْمَقَ. وَمَوْعِدُكَ أَيُّها الطِّفْلُ مَوْعِدُ نَباتٍ أَوْ طَيْرٍ.
> تُغْمِضُ عَيْنَيْكَ على ضُحًى تَتَساقَطُ مِنْ سِلالِهِ ٱلأَقْنِعَةُ، وَتَقْبُضُ بِكَفَّيْكَ على لِجامٍ غامِضٍ،
> كأَنَّما تَتَهَيَّأُ أَنْتَ لِلْكُهولةِ، أو تَتَهَيَّأُ لكَ ٱلكُهولةُ، لِتَخْتَزِلا مَعاً، ذلكَ السِّحْرَ ٱلَّذي يَنْبِضُ
> مَرَّةً واحِدَةً فَتَنْتَحِرُ ٱلحَياةُ شَوْقاً إلى نَبْضَةٍ ثانِيَةٍ.
>
> What is it that you see? Say, child, what is it that you see? Two hillocks on the horizon, a ribbon of villages, an earth staggering between reckless summer and witless winter. Your time, child, is like the time of plants or birds. You close your eyes upon a sunrise from whose baskets' disguises fall. You grasp a vague harness, as if preparing yourself for aging, or it preparing for you, so that you together may break the spell with only one pulsation, and then life destroys itself out of longing for a second.[55]

Salim Barakat famously writes in an archaic and difficult style with frequent unexpected word choices,[56] and many of the diacritics certainly facilitate reading comprehension by specifying the phonological form for rare words or inflectional forms. This, however, only motivates a small proportion of all the diacritics found in the text. The relative pronoun ٱلَّذي, for example, occurs three times in the excerpt, in all instances with the same diacritization. The word is extremely frequent, the 24th most common word in written Arabic,[57] and is therefore doubtlessly accessible via the direct route by the intended reader. It is not homographic with any other word and it does not have any variant non-standard vowel patterns that the diacritics may be intended to exclude. Similar analyses hold for the vast majority of words in the text. Case and noun inflection, such as the nominative inflection on the first noun in the excerpt, الطِّفْلُ 'the child,' could potentially serve to facilitate correct diction. This function is, however, superfluous (and possibly even inhibiting) in

[55] Salīm Barakāt, *al-Sīratān*, Beirut, Dār al-ǧadīd, 1998, p. 7. The translation presented here is indebted to the Swedish translation, Salim Barakat, *Järngräshoppan*, transl. Tetz Rooke, Stockholm, Tranan, 2000.

[56] Paul Starkey, *Modern Arabic Literature*, Edinburgh, Edinburgh University Press ("The New Edinburgh Islamic Surveys"), 2006, p. 153.

[57] Buckwalter and Parkinson, *A Frequency Dictionary of Arabic*, p. 11.

the form of private, silent reading in which novels are normally consumed, and could therefore also be interpreted as having an associative function in this text. A large proportion of diacritics throughout the novel, accordingly, are not motivated by factors relating to reading *per se*, but serve to associate the novel with a literary canon and to signal a claim to belonging to that canon or to lay claims to some of its qualities.

The second associative function is to signal that the text is to be associated, not necessarily with literary or linguistic qualities of culturally revered texts, but with the reading practices associated with them, in particular reading aloud, memorization, and recitation. One example of this is found in *al-Naḥw al-wāḍiḥ*,[58] a popular Arabic textbook for the 9th grade. Each chapter discusses a specific grammatical phenomenon and consists of four sections: examples (*al-amṯila*), discussion (*al-baḥṯ*), rules (*al-qāʿida*), and exercises (*tamrīn*). The examples below are from the chapter on indefinite nouns. In all chapters, the examples section (4a) and the rules section (4b) are presented with almost complete diacritization, with only *waṣla* missing, while the discussion section (4c) and the exercises (4d) are presented with only occasional diacritics.

(4)
a.

(١) فِي الدُّرْجِ كِتابٌ.
(٢) سَقَطَ مَنْزِلٌ فِي شَارِعِنا.

(1) There is a book in the drawer.
(2) A house in our street collapsed.

b.

(٦٩) اَلنَّكِرَةُ اسْم يَدُلُّ عَلَى شَيْءٍ غَيْر مُعَيَّن.
(٧٠) اَلمَعْرِفَةُ اسْم يَدُلُّ عَلَى شَيْءٍ مُعَيَّن.

(69) An indefinite noun refers to something non-specific.
(70) A definite noun refers to something specific.

[58] ʿAlī l-Ǧārim and Muṣṭafā Amīn, *al-Naḥw al-wāḍiḥ fī qawāʿid al-luġa l-ʿarabiyya li-madāris al-marḥala l-ūlā*, Cairo, Dār al-murtaḍā, 1983.

c.
إذا تأملنا كل اسم من الأسماء التي في الجمل السابقة، رأينا أن بعضها مثل: كتاب. ومنزل.
ورجل. وجواد. وتلميذ. وورقة، لا يدل على شيء معين معروف لنا [...]

If we reflect on each noun in the examples above, we see that some, like "book," "house," "man," "horse," "pupil," and "paper" do not refer to something specific known to us [...]

d.
اجعل المعرفة نكرة والنكرة معرفة في الجمل الآتية: (١) رَكب خادمٌ الحصان. (٢) طارت
ورقة من الكتاب.

Make the indefinite nouns definite and the definite nouns indefinite in the following sentence:
(1) A servant rode the horse. (2) A page flew away from the book.[59]

This difference in diacritization signals varying importance of the different sections and implies different intended reading practices. The complete diacritization in the examples and in the rules section implies that these are to be read aloud with attention to the exact linguistic form, and in the case of the rules section probably also that it is to be memorized. The lack of diacritics in the discussion implies that this section is not necessarily to be read aloud and that the focus is on the informational content, rather than on linguistic form.

5 Modes of Diacritization

Having now discussed types of diacritics and their potential functions, we turn to how they aggregate on the textual level. As mentioned in the introduction, the optionality of diacritics in theory allows for a practically infinite variability in how they are employed in a given text. In practice, however, forms of diacritization tend in modern typeset Standard Arabic to coalesce into a set of identifiable modes of diacritization. Seven such modes, described in more detail below, are identified here. They can be ordered according to the amount of diacritics, as follows:
1) no diacritization
2) disambiguating diacritization

[59] *Ibid.*, p. 195-196.

3) morphosyntactic diacritization
4) idiosyncratic diacritization
5) simplified diacritization
6) lexical diacritization
7) complete diacritization

As well as representing an increase in the proportion of diacritics added to the text, they are loosely structured in an implicational hierarchy of diacritization. This implicational hierarchy means that as the amount of diacritics in a text increases, types of diacritics tend to be added in a given order. The final establishment of this hierarchy requires detailed statistical investigation, which is beyond the scope of this article. Only a preliminary sketch is given here, with the hierarchy presented in (5) based on texts surveyed in the research for this article.

(5) disambiguating diacritics
> morphosyntactic diacritics; *šadda*
> short vowel diacritics
> long vowel diacritics; stem-internal *sukūn*
> *sukūn* after definite article; phonemically superfluous diacritics;
> *waṣla*; dagger *alif*

Any such hierarchy of diacritics must necessarily involve some fuzziness, since most texts with more than just minimally disambiguating diacritics display some measure of inconsistency. See for example (10) below, where the word جميل 'beautiful' is written with a long vowel diacritic ـِ on the first row but without it on the bottom row, or (8), where phonetically superfluous diacritics are added liberally, but seemingly at random. Rather, the hierarchy should be interpreted in terms of diacritics further down in the hierarchy being used in markedly lower proportions than those higher up. The hierarchy predicts, for example, that a text with all or most potential *šadda* will also include most or all disambiguating diacritics, or that a text that includes a large proportion of *waṣla*s will also include all or most morphosyntactic diacritics. Furthermore, the mode of lexical diacritization is a special case in that it clearly deviates from this hierarchy. It is also the least common of all seven modes.

Each mode is associated with an approximate rate of diacritization, defined as the percentage of all letters in a text that are followed by a diacritic. Since letters representing long vowels are never followed by diacritics, the maximum rate of diacritization is not 100%, but around 80%. The Qur'ān, for instance, the iconic representative of complete diacritization, has a rate of diacritization of 78.8%.[60] Some modes, and accordingly some rates of diacritization, are far more common than others. This is illustrated in figure 1, which shows the distribution of rates of diacritization in 1846 books of a variety of genres, published electronically by the Hindawi Foundation between 2008 and 2020.[61] The material includes new original works and translations, as well as new editions of Arabic books from throughout the 20th century which are now in the public domain. The books are divided here into normal prose (i.e. prose intended for skilled readers, *n* = 1646); children's literature (*n* = 137); and poetry (*n* = 73). As can be seen in figure 1, for normal prose there is a high concentration of rates slightly above zero, with a peak density of 1.2%, representing disambiguating diacritization. For children's literature, the peak density is 70.0%, representing simplified diacritization. For poetry the peak density is 9.5%. The rate for poetry is less informative than for the other genres, given the limited amount of data in this category. It is, however, surprisingly low, seeing as poetry is often described as being diacritized in the binary paradigm discussed in the introduction.

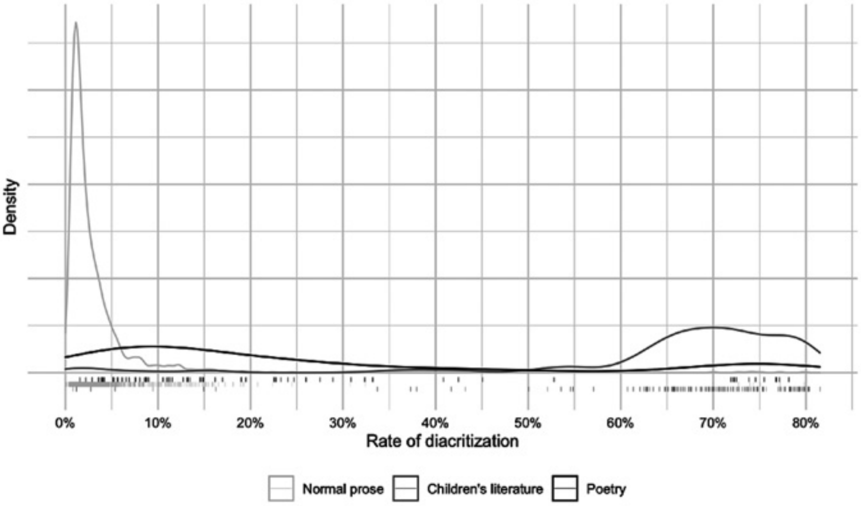

FIGURE 1 Density of rates of diacritization in Arabic books (n=1846). Vertical lines below the graph represent individual books in the three categories.

[60] This was calculated on the digitized Quranic text retrieved from http://tanzil.net on April 10, 2020.

[61] http://www.hindawi.org. Texts were retrieved on April 4, 2020.

In the remainder of this section, the seven modes of diacritization are discussed in turn.

Mode 1: No diacritization

Text completely void of diacritics is quite rare in book-length texts. As shown in figure 1, normal prose shows rates of diacritization with a narrow distribution near, but distinctly separate from, zero, with rates approaching zero quickly becoming increasingly rare.

Text with no diacritics is, however, common in digital non-standard vernacular Arabic text, which is not represented in the data in figure 1. The non-standard written vernacular is the unmarked form of writing in instant messaging services and social media.[62] It is also common in blog posts,[63] as well as in some literary genres.[64] The studies cited above feature numerous passages of non-standard Arabic written in the Arabic script, all completely devoid of diacritics. One of the reasons for the lack of diacritics in vernacular writing is certainly the inconvenience of typing them, especially given the speedy nature of text production characteristic of these media. A second reason is that there are simply fewer ambiguities in non-standard varieties of Arabic that need to be resolved by diacritics. Some of the common inflectional homographs of Standard Arabic, such as passives formed with changes in vowel patterns, are formed with consonantal affixes in the non-standard varieties[65] and therefore do not produce homographs. Furthermore,

[62] Madeline Haggan, "Text Messaging in Kuwait: Is the Medium the Message?," *Multilingua: Journal of Cross-Cultural and Interlanguage Communication*, 26/4 (2007), p. 427-449; Kristian Takvam Kindt, Jacob Høigilt and Tewodros Aragie Kebede, "Writing Change: Diglossia and Popular Writing Practices in Egypt," *Arabica*, 63/3 (2016), p. 324-376; Abdulrahman Alkhamees, Rasha Elabdali and Keith Walters, "Destabilizing Arabic Diglossia?," in *Perspectives on Arabic Linguistics XXXI: Papers from the Annual Symposium on Arabic Linguistics, Norman, Oklahoma, 2017*, eds Amel Khalfaoui and Youssef A. Haddad, Amsterdam, John Benjamins Publishing Company ("Studies in Arabic Linguistics," 8), 2019, p. 105-134.

[63] Teresa Pepe, *Fictionalized Identities in the Egyptian Blogosphere*, Oslo, University of Oslo, 2014; Dominique Caubet, "New Elaborate Written Forms of Darija: Blogging, Posting, and Slamming in Morocco," in *The Routledge Handbook of Arabic Linguistics*, eds Elabbas Benmamoun and Reem Bassiouney, London-New York, Routledge ("Routledge Language Handbooks"), 2018, p. 387-406.

[64] Eva Marie Håland, "Adab Sākhir (Satirical Literature) and the Use of Egyptian Vernacular," in *The Politics of Written Language in the Arab World*, eds Jacob Høigilt and Gunvor Mejdell, Leiden-Boston, Brill ("Studies in Semitic Languages and Linguistics," 90), 2017, p. 142-165.

[65] Jan Retsö, *The Finite Passive Voice in Modern Arabic Dialects*, Gothenburg, Acta Universitatis Gothoburgensis ("Orientalia Gothoburgensia," 7), 1983.

written Arabic vernacular varieties have developed conventions whereby frequent homographic words are disambiguated with letters. In Syrian Arabic, for example, the 2nd f. sg. pronoun *inti* is in writing differentiated from 2nd m. sg. *inta* in writing with a final letter ي, whereas in Standard Arabic the two analogous forms, *anta* and *anti*, can only be differentiated with diacritics.

Mode 2: Disambiguating diacritization

In the disambiguating diacritization mode, only those diacritics that facilitate word identification for proficient readers, as described above, are added to the text. This is by far the most common type of diacritization, being the unmarked mode for most types of text, including books, as shown in figure 1, where the diacritization corresponding to this mode has a peak density of 1.2%.

The systematic diacritization of all homographic words in a text would entail a substantial number of diacritics, giving rates far higher than 1.2%. Rather, in this mode diacritics are generally added to homographs only when the less frequent reading of a word is intended. For example, the letter sequence فِي corresponds to the words *fī* 'in,' *fayy* 'shadow,' and *fiyya* 'in me,' the first being extremely frequent and the latter two quite rare. In text with disambiguating diacritization, the word would only be supplied with diacritics if one of the two less frequent readings was intended. Similarly, most Arabic verbs are inflectional homographs with regard to voice, since voice is normally distinguished with short vowels (e.g. *kataba* 'wrote' vs. *kutiba* 'was written'). The active voice is the most frequent and is rarely disambiguated with diacritics, whereas the passive voice, which is less frequent, often is, unless the passive voice is clear from the immediate context.[66]

To further illustrate this mode, the first 50 pages of the novel *Šīkāgū*[67] were reviewed for words with diacritics. These pages include 29 words with one or more diacritics (excluding the nūnation diacritic ـً in connection with the letter *alif*, which is often systematically added in otherwise undiacritized text).[68] In all but four of these words, diacritics have a clear function of facilitating word identification, according to the subcategories of this function listed above. Of the 29 diacritized words, there are

– 13 lexical homographs (e.g. تَمَلَّكها 'took possession of her,' p. 18);
– 4 words with an opaque inflectional form (e.g. نوادٍ 'clubs,' p. 22);

[66] Hermena, Drieghe, Hellmuth and Liversedge, "Processing of Arabic Diacritical Marks," p. 596.

[67] Al-Aswānī, *Šīkāgū*.

[68] Hallberg, *Case Endings in Spoken Standard Arabic*, p. 77, n. 41.

- 4 inflectional homographs (all of them passives, e.g. يُفهم ‘was understood,’ p. 49);
- 3 rare words that the author may expect to be unfamiliar to some readers (e.g. شِبْشبًا ‘slippers,’ p. 19);
- 5 words that do not fit any of these categories (two instance of case inflection and three instance of dual forms with enclitic pronouns, e.g. كتفيْه ‘his shoulders,’ p. 25).

News prose employs a restricted form of disambiguating diacritization, with rates of diacritization above zero but below the rates common in book-length text. Table 1 lists rates of diacritization for all text from each of the five news agencies represented in the Arabic News Text Corpus.[69] The rate of diacritics in text from these sources ranges from 0.03% to 0.31%, well below the 1.2% typical of book-length normal prose shown above. These low rates may be due to the conventionalized, predictable style and restricted vocabulary of news prose,[70] limiting the need for disambiguation by means of diacritics. With these low rates of diacritics and the short format of news articles, a sizable proportion of news articles are completely void of diacritics, as indicated in the rightmost column in table 1. Indeed, averaging over these five news sources, half of all articles do not have a single diacritic.

Mode 3: Morphosyntactic diacritization

In the morphosyntactic diacritization mode, diacritics indicating case and mood inflection are systematically added, in addition to disambiguating

TABLE 1 Diacritization rates and article length in the Arabic News Text Corpus

Source	Articles	Mean article length (words)	Rate of dia.	Articles with no diacritics
Alarabia	6,519	284	0.31%	34%
BBC	3,815	417	0.11%	43%
CNN	3,507	235	0.24%	42%
France24	3,976	401	0.03%	81%
SkyNews	13,708	235	0.05%	76%
Mean		314	0.15%	55%

[69] Chouigui, Ben Khiroun and Elayeb, “An Arabic Multi-Source News Corpus.”

[70] Colleen Cotter, *News Talk: Investigating the Language of Journalism*, Cambridge-New York-Melbourne, Cambridge University Press, 2010, p. 136.

diacritics, giving a text where most diacritics appear in word final position. Because of the status of case and mood inflection as a marker of correct language, this mode features mainly in texts aimed at intermediate or advanced readers, such as in literature for adolescents, as in (6), and schoolbooks for secondary and tertiary education,[71] as in (7), taken from the Syrian government issued Arabic textbook for the 9th grade. These readers are assumed to have developed efficient word identification of undiacritized text and therefore do not require extensive word internal diacritics. They have, however, not developed skills in formal reading styles which require the word endings, and these endings are therefore supplied.

(6)

عندما وصلَ زياد إلى حاجزِ قلنديا، كانتْ أشعّةُ شمسِ الصّباح تزيحُ عتمةَ اللّيلِ بخجلٍ
مظهرةً طابورَ سيّاراتٍ ملتويًا كأفعى طويلةٍ على مدى البصرِ. كانَ زياد يستيقظُ كلَّ يومٍ
معَ أذانِ الفجرِ، فيخرجُ منَ البيتِ في محاولةٍ منهُ للوصولِ باكرًا إلى الحاجزِ.

When Ziyād arrived at the Qalandiyā checkpoint, the morning sun was already shyly swiping away the darkness of the night, revealing the queue of cars winding like a long snake as far as you could see. Every morning, Ziyād woke up to the call for prayer and left the house to get to the checkpoint as early as possible.[72]

(7)

* أقرأُ النصَّ الآتي:

تزخرُ أرضُ سوريَةَ بالكثيرِ من الآثارِ التي تشهدُ على عراقةِ هذا البلدِ ومكانتِهِ التَّاريخيَّةِ
المرموقةِ، وتحكي قصَّةَ جمالِ الفَنِّ وسحرِهِ، ومن يشاهدُها يعجبُ بها إعجاباً، ويقدِّرُ حضارةَ
أهلِها تقديراً كثيراً، فيزورُها السُّيَّاحُ من كلِّ مكانٍ، ويندهشونَ بروعتِها كلَّ الاندهاشِ،
فهم يجولونَ فيها جولتينِ أو أكثر مأخوذينَ ببراعةِ مَن أشادَها [...]

* I read the following text:

[71] Muhammad al-Sharkawi, "The Ecology of Case in Modern Standard Arabic," *Folia Orientalia*, 53 (2016), p. 249; Al Midhwah and Alhawary, "Arabic Diacritics and Their Role in Facilitating Reading Speed, Accuracy, and Comprehension by English L2 Learners of Arabic," p. 24-26.

[72] Taġrīd al-Naǧǧār and Ǧulnār Ḥāǧū, *Sitt al-kull*, Amman, Salwa Publishers, 2015, p. 1.

> The Syrian lands are rich with historical sites that bear witness to the heritage of this land and its historical importance. They tell the story of the beauty and the mystery of the arts, and whoever beholds them is amazed and appreciative of the culture of their people. Tourists from all over the world come to visit and are astonished by their magnificence, traveling the lands several times over, taken by the creativity of their founders […][73]

Morphosyntactic diacritization, together with the idiosyncratic diacritization described below, can thus be characterized as "half-diacritized" text, falling into the middle range of rates of diacritization. The relative rarity of these two modes can be clearly seen in figure 1 in the scarcity of books with rates of diacritization between 20% and 60%.

Mode 4: Idiosyncratic diacritization

Idiosyncratic diacritization is here adopted as a catch-all mode for patterns of diacritization with significantly more diacritics than is required for disambiguation, and significantly less than in the simplified diacritization mode discussed below, but where diacritization does not follow a discernible pattern. It overlaps with morphosyntactic diacritization in terms of its rate of diacritization in the middle range, but differs from it in not following a pattern according to which diacritics are added. An example of this mode is (8), taken from an Arabic translation of a scholarly work on pre-Islamic Arabs.

(8)
يظهرُ أناسٌ يحملُون اسمَ الـ ’:عرَب‘ [كذا] لأول مرةٍ في مَصَادِرَ ترتبطُ بأحْدَاث في
سُورِيَا في القُرُون الأولى من الألفِ الأخِيرَة قَبلَ المِيلاد. وأولُ هذه المَصَادِر هي
النُّصُوصُ الأكَّادِيَّةُ التي وصَلَت إلينا من آشُور: سجِلاتُ المُلُوك التي يتحدثُون فيها عن
حَمْلاتِهم إلى سُورِيَا والصَّحْرَاءِ السُّورِيَّة منذ منتصَفِ القَرنِ التاسع قَبلَ المِيلادِ حتَّى
سقُوط الإمبْرَاطُورِيَّةِ.

> People designated as Arabs first appear in sources connected with events in Syria in the first centuries of the first millennium BC. The main ones are the texts in Akkadian from Assyria: the records of kings telling about

[73] Niḍāl al-Ṣāliḥ (ed.), *al-Luġa l-ʿarabiyya: al-Ṣaff al-tāsiʿ al-asāsī*, Damascus, National Center for Curriculum Development, 2019, p. 22.

> their campaigns to Syria and the Syrian desert from the middle of the ninth century down to the fall of the empire.[74]

To the extent that diacritization in this mode extends beyond disambiguation, its function is best interpreted as associative. The high rates of diacritization signal care for correct language and respect for the linguistic and literary heritage, and possibly also that the text is difficult and to be taken seriously. By not extending diacritization to the mode of simplified diacritization, nor employing morphosyntactic diacritization, the association with schoolbooks is avoided.

Mode 5: Lexical diacritization

In the lexical diacritization mode, word internal diacritics are systematically added and word final morphosyntactic diacritics systematically omitted, giving, in effect, the inverse of morphosyntactic diacritization. This mode makes the indirect route available for every word, to the benefit of novice readers, while avoiding the representation of many features specific to very formal styles of diction. As such, it closely and transparently represents word forms as phonemically encoded in silent reading, as explained above. For these reasons, some researchers have adopted it for stimulus texts in studies on reading development.[75]

Outside of experimental settings, however, lexical diacritization is virtually unheard of. The only example found in the research for this article is *al-Kitāb al-asāsī*,[76] an entry level Arabic textbook for second language learners. The authors explain in the introduction that the morphosyntactic endings are too difficult for the beginner level.[77] This mode is completely consistent in the first 13 of the 25 chapters, as exemplified in (9). This quote has a rate of diacritization of 63%, which is similar to the rates of the more common simplified diacritization discussed below, but differs from it in its distribution of diacritics. The remaining chapters in the book employ the simplified diacritization

[74] Jan Retsö, *al-ʿArab fī l-ʿuṣūr al-qadīma min al-āšūriyyīn ilā l-umawiyyīn*, Riad, King Saud University Press, 2016, I, p. 161. English original in Jan Retsö, *The Arabs in Antiquity: Their History from the Assyrians to the Umayyads*, London-New York, RoutledgeCurzon, 2005, p. 119.

[75] Saiegh-Haddad and Schiff, "The Impact of Diglossia on Voweled and Unvoweled Word Reading in Arabic"; Saiegh-Haddad, "MAWRID."

[76] Al-Saʿīd M. Badawī and Fatḥī ʿAlī Yūnus, *al-Kitāb al-asāsī fī taʿlīm al-luġa l-ʿarabiyya li-ġayr al-nāṭiqīn bi-hā*, Tunis, al-Munaẓẓama l-ʿarabiyya li-l-tarbiya wa-l-ṯaqāfa wa-l-ʿulūm, 1973.

[77] *Ibid.*, p. ح [viii].

mode, which includes case and mood inflection, in accordance with the traditional view that even relatively novice readers need to master this system.

(9)

أَحْمَد ـ الزَّوْج ـ مُهَنْدِس، فَاطِمَة ـ الزَّوْجَة ـ مُوَظَّفَة فِى مَكْتَب الْبَرِيد. لِأَحْمَد وَزَوْجَته أَرْبَعَة
أَوْلَاد: وَلَدَان وَبِنْتَان. مُحَمَّد فِى السَّادِسَة عَشْرَة ، وَهُوَ تِلْمِيذ فِى الْمَدْرَسَة الثَّانَوِيَّة، وَمَحْمُود ـ
أَخُوه ـ فِى الثَّالِثَة عَشْرَة ، وَهُوَ تِلْمِيذ فِى الْمَدْرَسَة الأِعْدَادِيَّة [...]

Aḥmad, the husband, is an engineer. Fāṭima, the wife, is an employee at the post office. Aḥmad and his wife have four children: two boys and two girls. Muḥammad is sixteen years old and is a high-school student. Maḥmūd, his brother, is thirteen, and he is a student in primary school [...][78]

Mode 6: Simplified diacritization

The simplified diacritization mode includes all potential diacritics, except, typically, those in the last two steps in the hierarchy of diacritics in (5) (*sukūn* in connection with definite article, phonemically superfluous diacritics, *waṣla*, and dagger *alif*), resulting in a diacritization rate of roughly 60-70%. This is the go-to style of diacritization when the intention is to produce what is commonly identified as "vowelled" text, for example in children's literature and Arabic primers. The latter is exemplified in (10), taken from the Egyptian government issued first grade Arabic textbook. The dominance of this mode in children's literature is clearly seen in figure 1.

(10)

أَسْكُنُ فى شَارِعٍ جَمِيلٍ.
أَتَعاونُ مَعَ الجيرانِ.
نَزْرَعُ الأَشْجَارَ.
نُزَيِّنُ الجدْرانَ.
ونُحافِظُ عَلَى الأزْهَارِ.
وَنضَعُ القُمَامةَ فى صُنْدُوقِ القُمَامةِ.

[78] *Ibid.*, p. 166.

شَارِعُنَا نَظِيفٌ.
شَارِعُنَا جَمِيلٌ.

> I live in a beautiful street.
> I cooperate with my neighbors.
> We plant trees.
> We decorate the walls
> and take care of the flowers.
> We throw the trash in the trash bin.
> Our street is clean.
> Our street is beautiful.[79]

This mode facilitates reading comprehension for novice readers by making the indirect route (assembled phonology) available for all words. As opposed to the lexical diacritization mode, however, it also heavily features diacritics favoring correct diction, most importantly case and mood inflection, which is not representative of skilled silent reading and which may negatively affect word identification.

The use of this mode in texts intended for skilled readers, on the other hand, is a highly marked choice, and the function of diacritics in those texts is primarily associative. In the corpus on which figure 1 is based, only seven of the 1646 books of normal prose have a diacritics rate above 50%. Example (3), discussed above, with a diacritization rate of 63%, is an example of this.

Mode 7: Complete diacritization

Texts written in the complete diacritization mode features complete and consistent use of all potential diacritics, with the possible exception of dagger *alif* and *waṣla*, which are often omitted also in this mode. It is represented at the right edge in figure 1 with a diacritization rate of approximately 75-80%. The iconic text for this mode is the Qurʾān, and it is also used in Arabic translations of the Bible, as shown in (11). This mode carries prestige by virtue of association with the Qurʾān, the ideological model of good Arabic,[80] as well as other religious texts. It features a large number of diacritics that are of little benefit to the reader, even for correct diction, such as phonemically superfluous

[79] Farağ and Ġurāb, *Hayyā natawāṣal*, p. 2.

[80] Bohas, Guillaume and Kouloughli, *The Arabic Linguistic Tradition*, p. 18; Yasir Suleiman, "The Simplification of Arabic Grammar and the Problematic Nature of the Sources," *Journal of Semitic Studies*, 41/1 (1996), p. 114.

diacritics and word final diacritics in pause position, the latter being prescriptively not pronounced.

(11)

١فِي ٱلْبَدْءِ كَانَ ٱلْكَلِمَةُ، وَٱلْكَلِمَةُ كَانَ عِنْدَ ٱللّٰهِ، وَكَانَ ٱلْكَلِمَةُ ٱللّٰهَ. ٢هٰذَا كَانَ فِي ٱلْبَدْءِ عِنْدَ
ٱللّٰهِ. ٣كُلُّ شَيْءٍ بِهِ كَانَ، وَبِغَيْرِهِ لَمْ يَكُنْ شَيْءٌ مِمَّا كَانَ.

In the beginning was the Word, and the Word was with God, and the Word was God. He was with God in the beginning. Through him all things were made; without him nothing was made that has been made.[81]

The complete diacritization mode is also used for children's literature, although to a lesser extent than simplified diacritization. This is visible in figure 1 at the right edge in the line for children's literature. Of the 137 children's books in the corpus, 29 have a diacritization rate above 75%, which is characteristic of this mode. These books feature all potential diacritics except for *waṣla* and dagger *alif*, and substitute ـَ for dagger *alif*. The first lines of one of these books are shown in (12).

(12)

- مَرْحَبًا بِكَ يَا « سَعِيدُ »!
شَدَّ مَا آنَسْتَنَا وَمَلَأْتَ قُلُوبَنَا فَرَحًا بِقُدُومِكَ.
- شُكْرًا لَكَ يَا « صَلَاحُ »، فَإِنَّ فَرَحِي بِلِقَائِكَ لَا يُوصَفُ ... وَلَسْتُ أُغَالِي إِذَا قُلْتُ لَكَ
إِنَّنِي كُنْتُ أَعُدُّ الْأَيَّامَ بِفَارِغِ الصَّبْرِ، لِأَقْضِيَ مَعَكُمْ إِجَازَةَ هَذَا الْعَامِ، كَمَا قَضَيْتُ إِجَازَةَ
الْعَامِ السَّابِقِ.

– Hello, Saʿīd!
How we have missed you and are happy now that you have arrived!
– Thank you, Selim. I am indescribably happy to see you, and I am not exaggerating when I say that I have impotently been counting the days until I will get to spend the holidays with you this year, as we did last year.[82]

[81] Cornelius Van Alen Van Dyck (transl.), *al-Kitāb al-Muqaddas*, Cairo, Bible Society of Egypt, 1999, p. 120. English translation from *New International Version*.

[82] Kāmil Kaylānī, *Bayna ʿaṣr al-ẓalām wa-maṭlaʿ al-faǧr*, Windsor, Hindawi, 2019, p. 6.

Conclusion

Arabic diacritics are in modern typeset text used to different extents, in different patterns, and serve a variety of different functions. This article has presented a framework for conceptualizing and describing this variation. On the sub-word level, diacritics can be categorized in terms of their interaction with the letters, as phonetically additive, specifying, or superfluous. With regard to the reading process, diacritics are suggested to serve three main functions: to facilitate reading comprehension, to facilitate prescriptively correct diction, and to evoke associations with other texts. Differences in diacritization between texts and genres can be characterized according to the relative emphasis given to these three functions. A classification of patterns of diacritization into seven "modes," ranging from no diacritization to complete diacritization, has been presented, providing a convenient way of describing various forms of partial diacritization. Furthermore, each mode is closely, but not absolutely, associated with a set of genres.

This view of Arabic diacritization as a complex, multi-faceted phenomenon suggests several lines of further research. The hierarchically governed system of diacritics, a preliminary sketch of which was presented above, could be quantitatively tested and described. This would require a large corpus of varied texts and a method for computationally identifying diacritics in various positions. Qualitative and detailed investigations and comparisons of forms of diacritization in individual works could also yield interesting results. Investigating the views of text producers, authors, and publishers, could also potentially help explain the sometimes puzzling variation in the use of diacritics described here.

Furthermore, the view of diacritization as multi-functional has far-reaching implications for research on the effects of diacritics on reading. This research has almost exclusively been based on a binary concept of diacritization, comparing the reading of text with either no diacritization or complete diacritization, and has produced mixed results regarding its effects on comprehension and reading-aloud accuracy.[83] Teasing out the contribution to reading of

[83] Bashir Abu-Hamour, Hanan Al-Hmouz and Mohammed Kenana, "The Effect of Short Vowelization on Curriculum-Based Measurement of Reading Fluency and Comprehension in Arabic," *Australian Journal of Learning Difficulties*, 18/2 (2013), p. 181-197, and Salim Abu-Rabia and Abedalhakeem Salfeety, "Reading in Arabic Orthography: The Influence of Short Vowels on Reading Accuracy and Comprehension of Poor and Normal Arabic Readers," *Journal of Advances in Linguistics*, 5/2 (2015), p. 723-735, for example, report positive effects for diacritics on these measures, while no or negative effects are reported in Abdullah M. Seraye, "Short Vowels Versus Word Familiarity in the Reading Comprehension of

different sets of diacritics with different functions (facilitating word identification and correct diction, for example) may help explain these mixed results and could result in evidence-based recommendations for the partial use of diacritics for improved readability for different types of target readers.

Bibliography

Corpus

ʿAbd al-Ġanī, Ayman Amīn, *al-Kāfī fī qawāʿid al-imlāʾ wa-l-kitāba*, Cairo, Dār al-tawfīqiyya li-l-turāṯ, 2012.

Al-Aswānī, ʿAlāʾ, *Šīkāgū*, Cairo, Dār al-šurūq, 2007.

Badawī, al-Saʿīd M., and Fatḥī ʿAlī Yūnus, *al-Kitāb al-asāsī fī taʿlīm al-luġa l-ʿarabiyya li-ġayr al-nāṭiqīn bi-hā*, Tunis, al-Munaẓẓama l-ʿarabiyya li-l-tarbiya wa-l-ṯaqāfa wa-l-ʿulūm, 1973.

Barakāt, Salīm, *al-Sīratān*, Beirut, Dār al-ǧadīd, 1998.

Barakāt, Salīm, *Järngräshoppan*, transl. Tetz Rooke, Stockholm, Tranan, 2000.

Farağ, Muḥammad Ṣalāḥ, and Muḥammad ʿAbd al-Ḥamīd Ġurāb, *Hayyā natawāṣal: al-luġa l-ʿarabiyya li-l-ṣaff al-awwal al-ibtidāʾī, al-faṣl al-dirāsī l-ṯānī*, Cairo, Wizārat al-tarbiya wa-l-taʿlīm, 2008.

Al-Ǧārim, ʿAlī, and Muṣṭafā Amīn, *al-Naḥw al-wāḍiḥ fī qawāʿid al-luġa l-ʿarabiyya li-madāris al-marḥala l-ūlā*, Cairo, Dār al-murtaḍā, 1983.

Idrīs, Yūsuf, *al-Ab al-ġāʾib*, Cairo, Maktabat Miṣr, 1987.

Idrīs, Yūsuf, *al-Ab al-ġāʾib*, Windsor, Hindawi, 2017.

Kaylānī, Kāmil, *Bayna ʿaṣr al-ẓalām wa-maṭlaʿ al-fağr*, Windsor, Hindawi, 2019.

Al-Naǧǧār, Taġrīd, and Ǧulnār Ḥāǧū, *Sitt al-kull*, Amman, Salwa Publishers, 2015.

Retsö, Jan, *The Arabs in Antiquity: Their History from the Assyrians to the Umayyads*, London-New York, RoutledgeCurzon, 2005.

Retsö, Jan, *al-ʿArab fī l-ʿuṣūr al-qadīma min al-āšūriyyīn ilā l-umawiyyīn*, Riad, King Saud University Press, 2016.

Al-Ṣāliḥ, Niḍāl (ed.), *al-Luġa l-ʿarabiyya: al-Ṣaff al-tāsiʿ al-asāsī*, Damascus, National Center for Curriculum Development, 2019.

Van Alen Van Dyck, Cornelius (transl.), *al-Kitāb al-Muqaddas*, Cairo, Bible Society of Egypt, 1999.

Arab Readers: A Revisited Issue," *International Electronic Journal of Elementary Education*, 8/3 (2016), p. 481-506; Ibrahim A. Asadi, "Reading Arabic with the Diacritics for Short Vowels: Vowelised but Not Necessarily Easy to Read," *Writing Systems Research*, 9/2 (2017), p. 137-147; and Taha, "Deep and shallow in Arabic orthography."

Studies

Abu-Hamour, Bashir, Hanan Al-Hmouz and Mohammed Kenana, "The Effect of Short Vowelization on Curriculum-Based Measurement of Reading Fluency and Comprehension in Arabic," *Australian Journal of Learning Difficulties*, 18/2 (2013), p. 181-197.

Abu-Leil, Aula Khateeb, David L. Share and Raphiq Ibrahim, "How Does Speed and Accuracy in Reading Relate to Reading Comprehension in Arabic?," *Psicologica: International Journal of Methodology and Experimental Psychology*, 35/2 (2014), p. 251-276.

Abu-Rabia, Salim, "Reading Arabic Texts: Effects of Text Type, Reader Type and Vowelization," *Reading and Writing*, 10/2 (1998), p. 105-119.

Abu-Rabia, Salim, and Abedalhakeem Salfeety, "Reading in Arabic Orthography: The Influence of Short Vowels on Reading Accuracy and Comprehension of Poor and Normal Arabic Readers," *Journal of Advances in Linguistics*, 5/2 (2015), p. 723-735.

Al Midhwah, Ali, and Mohammad T. Alhawary, "Arabic Diacritics and Their Role in Facilitating Reading Speed, Accuracy, and Comprehension by English L2 Learners of Arabic," *The Modern Language Journal*, 104/2 (2020), p. 418-438.

Alkhamees, Abdulrahman, Rasha Elabdali and Keith Walters, "Destabilizing Arabic Diglossia?," in *Perspectives on Arabic Linguistics XXXI: Papers from the Annual Symposium on Arabic Linguistics, Norman, Oklahoma, 2017*, eds Amel Khalfaoui and Youssef A. Haddad, Amsterdam, John Benjamins Publishing Company ("Studies in Arabic Linguistics," 8), 2019, p. 105-134.

Arts, Tressy, Yonatan Belinkov, Nizar Habash, Adam Kilgarriff and Vit Suchomel, "arTenTen: Arabic Corpus and Word Sketches," *Journal of King Saud University – Computer and Information Sciences*, 26/4 (2014), p. 357-371.

Asadi, Ibrahim A., "Reading Arabic with the Diacritics for Short Vowels: Vowelised but Not Necessarily Easy to Read," *Writing Systems Research*, 9/2 (2017), p. 137-147.

Badawi, El-Said M., Michael G. Carter and Adrian Gully, *Modern Written Arabic: A Comprehensive Grammar*, London-New York, Routledge ("Routledge Comprehensive Grammars"), 2004.

Bateson, Mary Catherine, *Arabic Language Handbook*, Washington, Center for Applied Linguistics ("Language Handbook Series"), 1967.

Bohas, Georges, Jean-Patrick Guillaume and Djamel Eddine Kouloughli, *The Arabic Linguistic Tradition*, London-New York, Routledge ("Arabic Thought and Culture"), 1990.

Buckwalter, Timothy, "Issues in Arabic Morphological Analysis," in *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, eds Abdelhadi Soudi, Antal van den Bosch and Günter Neumann, Dordrecht, Springer ("Text, Speech and Language Technology," 38), 2007, p. 23-41.

Buckwalter, Tim, and Dilworth B. Parkinson, *A Frequency Dictionary of Arabic: Core Vocabulary for Learners*, London-New York, Routledge ("Routledge Frequency Dictionaries"), 2011.

Caubet, Dominique, "New Elaborate Written Forms of Darija: Blogging, Posting, and Slamming in Morocco," in *The Routledge Handbook of Arabic Linguistics*, eds Elabbas Benmamoun and Reem Bassiouney, London-New York, Routledge ("Routledge Language Handbooks"), 2018, p. 387-406.

Chouigui, Amina, Oussama Ben Khiroun and Bilel Elayeb, "An Arabic Multi-Source News Corpus: Experimenting on Single-Document Extractive Summarization," *Arabian Journal for Science and Engineering*, 46/4 (2021), p. 3925-3938.

Coltheart, Max, "Dual Route and Connectionist Models of Reading: An Overview," *London Review of Education*, 4/1 (2006), p. 5-17.

Coltheart, Max, Kathleen Rastle, Conrad Perry, Robyn Langdon and Johannes C. Ziegler, "DRC: A Dual Route Cascaded Model of Visual Word Recognition and Reading Aloud," *Psychological Review*, 108/1 (2001), p. 204-256.

Cotter, Colleen, *News Talk: Investigating the Language of Journalism*, Cambridge-New York-Melbourne, Cambridge University Press, 2010.

Al-Daḥdāḥ, Anṭwān, *Muʿǧam qawāʿid al-luġa l-ʿarabiyya fī ǧadāwil wa-lawḥāt*, Beirut, Maktabat Lubnān, 1992.

Daniels, Peter T., "Fundamentals of Grammatology," *Journal of the American Oriental Society*, 110/4 (1990), p. 727-731.

Daniels, Peter T., "The Arabic Writing System," in *The Oxford Handbook of Arabic Linguistics*, ed. Jonathan Owens, Oxford-New York, Oxford University Press ("Oxford Handbooks in Linguistics"), 2013, p. 212-232.

Ferguson, Charles A., "Diglossia," *Word*, 15/2 (1959), p. 325-340.

Ferguson, Charles A., "Epilogue: Diglossia Revisited," in *Understanding Arabic: Essays in Contemporary Arabic Linguistics in Honor of El-Said Badawi*, ed. Alaa Elgibali, Cairo, American University in Cairo Press, 1996, p. 49-67.

Habash, Nizar Y., *Introduction to Arabic Natural Language Processing*, San Rafael, Morgan & Claywood ("Synthesis Lectures on Human Language Technologies," 10), 2010.

Haeri, Niloofar, *Sacred Language, Ordinary People: Dilemmas of Culture and Politics in Egypt*, New York-Basingstoke, Palgrave Macmillan, 2003.

Haggan, Madeline, "Text Messaging in Kuwait: Is the Medium the Message?," *Multilingua: Journal of Cross-Cultural and Interlanguage Communication*, 26/4 (2007), p. 427-449.

Håland, Eva Marie, "Adab Sākhir (Satirical Literature) and the Use of Egyptian Vernacular," in *The Politics of Written Language in the Arab World*, eds Jacob Høigilt and Gunvor Mejdell, Leiden-Boston, Brill ("Studies in Semitic Languages and Linguistics," 90), 2017, p. 142-165.

Hallberg, Andreas, *Case Endings in Spoken Standard Arabic: Statistics, Norms, and Diversity in Unscripted Formal Speech*, Doctoral dissertation, Lund University, 2016.

Hallberg, Andreas, and Diederick C. Niehorster, "Parsing Written Language with Non-Standard Grammar: An Eye-Tracking Study of Case Marking in Arabic," *Reading and Writing*, 34/1 (2021), p. 27-48.

Harris, Margaret, and Max Coltheart, *Language Processing in Children and Adults: An Introduction*, London-New York, Routledge ("Introductions to Modern Psychology"), 1989.

Hermena, Ehab W., Denis Drieghe, Sam Hellmuth and Simon P. Liversedge, "Processing of Arabic Diacritical Marks: Phonological/Syntactic Disambiguation of Homographic Verbs and Visual Crowding Effects," *Journal of Experimental Psychology: Human Perception and Performance*, 41/2 (2015), p. 494-507.

Holes, Clive, *Modern Arabic: Structures, Functions, and Varieties*, Washington, Georgetown University Press ("Georgetown Classics in Arabic Language and Linguistics"), 2004.

Ibn ʿAqīl, ʿAbd Allāh, *Šarḥ Ibn ʿAqīl ʿalā Alfiyyat Ibn Mālik*, ed. Anṭūniyūs Buḍrus, Tripoli, al-Muʾassasa l-ḥadīṯa li-l-kitāb, 2005.

Ibrahim, Muhammad H., "Linguistic Distance and Literacy in Arabic," *Journal of Pragmatics*, 7/5 (1983), p. 507-515.

Ibrahim, Raphiq, "Reading in Arabic: New Evidence for the Role of Vowel Signs," *Creative Education*, 4/4 (2013), p. 248-253.

Kessler, Brett, and Rebecca Treiman, "Writing Systems: Their Properties and Implications for Reading," in *The Oxford Handbook of Reading*, eds Alexander Pollatsek and Rebecca Treiman, New York, Oxford University Press ("Oxford Library of Psychology"), 2015, p. 10-25.

Kindt, Kristian Takvam, Jacob Høigilt and Tewodros Aragie Kebede, "Writing Change: Diglossia and Popular Writing Practices in Egypt," *Arabica*, 63/3 (2016), p. 324-376.

Maamouri, Mohamed, *Language Education and Human Development: Arabic Diglossia and Its Impact on the Quality of Education in the Arab Region*, discussion paper prepared for The World Bank, The Mediterranean Development Forum, Marrakech, 3-6 September 1998, Philadelphia, International Literacy Institute, 1998.

Meletis, Dimitrios, "The Grapheme as a Universal Basic Unit of Writing," *Writing Systems Research*, 11/1 (2019), p. 26-49.

Nelson, Kristina, *The Art of Reciting the Qurʾan*, Cairo-New York, The American University in Cairo Press, 2001.

Nemeth, Titus, "Simplified Arabic: A New Form of Arabic Type for Hot-Metal Composition," in *Typography Papers 9*, eds Eric Kindel and Paul Luna, London, Hyphen, 2013, p. 173-189.

Nemeth, Titus, "Arabic Hot Metal," *Philological Encounters*, 3/4 (2018), p. 496-523.

Parkinson, Dilworth B., “Searching for Modern Fuṣḥa: Real-Life Formal Arabic,” *al-ʿArabiyya*, 24 (1991), p. 31-64.

Parkinson, Dilworth B., “Under the Hood of arabiCorpus,” in *Arabic Corpus Linguistics*, eds Tony McEnery, Andrew Hardie and Nagawa Younis, Edinburgh, Edinburgh University Press, 2019, p. 17-29.

Pepe, Teresa, *Fictionalized Identities in the Egyptian Blogosphere*, Oslo, University of Oslo, 2014.

Pollatsek, Alexander, “The Role of Sound in Silent Reading,” in *The Oxford Handbook of Reading*, eds Alexander Pollatsek and Rebecca Treiman, New York, Oxford University Press (“Oxford Library of Psychology”), 2015, p. 185-201.

Rayner, Keith, Alexander Pollatsek, Jane Ashby and Charles Clifton, *Psychology of Reading*, New York-London, Prentice Hall, 2012.

Retsö, Jan, *The Finite Passive Voice in Modern Arabic Dialects*, Gothenburg, Acta Universitatis Gothoburgensis (“Orientalia Gothoburgensia,” 7), 1983.

Rogers, Henry, *Writing Systems: A Linguistic Approach*, Malden, Blackwell Publishers (“Blackwell Textbooks in Linguistics,” 18), 2005.

Roman, G., and B. Pavard, “A Comparative Study: How We Read in Arabic and French,” in *Eye Movements from Physiology to Cognition*, eds John Kevin O’Regan and Ariane Levy-Schoen, Amsterdam-New York, North-Holland, 1987, p. 431-440.

Ryding, Karin C., *A Reference Grammar of Modern Standard Arabic*, Cambridge-New York-Melbourne, Cambridge University Press, 2005.

Saiegh-Haddad, Elinor, “MAWRID: A Model of Arabic Word Reading in Development,” *Journal of Learning Disabilities*, 51/5 (2018), p. 454-462.

Saiegh-Haddad, Elinor, and Roni Henkin-Roitfarb, “The Structure of Arabic Language and Orthography,” in *Handbook of Arabic Literacy: Insights and Perspectives*, eds Elinor Saiegh-Haddad and R. Malatesha Joshi, New York, Springer (“Literacy Studies,” 9), 2014, p. 3-28.

Saiegh-Haddad, Elinor, and Rachel Schiff, “The Impact of Diglossia on Voweled and Unvoweled Word Reading in Arabic: A Developmental Study from Childhood to Adolescence,” *Scientific Studies of Reading*, 20/4 (2016), p. 311-324.

Saiegh-Haddad, Elinor, and Bernard Spolsky, “Acquiring Literacy in a Diglossic Context: Problems and Prospects,” in *Handbook of Arabic Literacy: Insights and Perspectives*, eds Elinor Saiegh-Haddad and R. Malatesha Joshi, Dordrecht, Springer (“Literacy Studies,” 9), 2014, p. 225-240.

Schulz, Eckehard, Günther Krahl and Wolfgang Reuschel, *Standard Arabic: An Elementary/Intermediate Course*, Cambridge, Cambridge University Press, 2000.

Seraye, Abdullah M., “Short Vowels Versus Word Familiarity in the Reading Comprehension of Arab Readers: A Revisited Issue,” *International Electronic Journal of Elementary Education*, 8/3 (2016), p. 481-506.

Share, David L., "On the Anglocentricities of Current Reading Research and Practice: The Perils of Overreliance on an 'Outlier' Orthography," *Psychological Bulletin*, 134/4 (2008), p. 584-615.

Share, David L., and Amalia Bar-On, "Learning to Read a Semitic Abjad: The Triplex Model of Hebrew Reading Development," *Journal of Learning Disabilities*, 51/5 (2018), p. 444-453.

Al-Sharkawi, Muhammad, "The Ecology of Case in Modern Standard Arabic," *Folia Orientalia*, 53 (2016), p. 223-259.

Starkey, Paul, *Modern Arabic Literature*, Edinburgh, Edinburgh University Press ("The New Edinburgh Islamic Surveys"), 2006.

Stetkevych, Jaroslav, *The Modern Arabic Literary Language: Lexical and Stylistic Developments*, Washington, Georgetown University Press ("Georgetown Classics in Arabic Language and Linguistics"), 2006.

Suleiman, Yasir, "The Simplification of Arabic Grammar and the Problematic Nature of the Sources," *Journal of Semitic Studies*, 41/1 (1996), p. 99-119.

Taha, Haitham, "Deep and Shallow in Arabic Orthography: New Evidence from Reading Performance of Elementary School Native Arab Readers," *Writing Systems Research*, 8/2 (2016), p. 133-142.

Taouka, Miriam, and Max Coltheart, "The Cognitive Processes Involved in Learning to Read in Arabic," *Reading and Writing*, 17/1 (2004), p. 27-57.

The Unicode Standard: Version 12.0 – Core Specifications, Mountain View, Unicode Consortium, 2019.

Uhlmann, Allon J., "Arabs and Arabic Grammar Instruction in Israeli Universities: Alterity, Alienation and Dislocation," *Middle East Critique*, 21/1 (2012), p. 101-116.

Versteegh, Kees, "Arabic Grammar and Corruption of Speech," *al-Abḥāṯ*, 31 (1983), p. 139-160.

Versteegh, Kees, "Learning Arabic in the Islamic World," in *The Foundations of Arabic Linguistics III: The Development of a Tradition: Continuity and Change*, eds Georgine Ayoub and Kees Versteegh, Leiden-Boston, Brill ("Studies in Semitic Languages and Linguistics," 94), 2018, p. 245-267.

Wagner, Daniel A., and Abdelhamid Lotfi, "Learning to Read by 'Rote,'" *International Journal of the Sociology of Language*, 42 (1983), p. 111-121.