Word Reading in Arabic: Influences of Diacritics and Ambiguity

Michael Grosvald, Sarah Al-Alami, and Ali Idrissi

1. Introduction

In the Arabic orthographic system, consonants and long vowels are written, while short vowel diacritics are generally omitted. Although this situation is well-known, the specific role that diacritics play in the processing of written Arabic words by adult readers remains understudied. Reading studies incorporating tasks such as lexical decision are well-established in the literature regarding other languages such as English, but relatively few such studies have been carried out on Arabic word reading.

The Arabic orthographic system has three diacritic symbols which can be used to represent the short vowel phonemes /u/, /i/ and /a/ (respectively: ـُـ ـِـ ـَـ, with the dashes indicating the consonant symbol above or below which the diacritic would appear). A vowel diacritic appearing above or below a consonant symbol indicates that as pronounced, the consonant is followed by the short vowel. Other diacritics are used in order to indicate the absence of such a short vowel (as when a syllable has a non-empty coda) or word-final case marker (addition of a short vowel plus a final /n/: anـًـ, inـٍـ and unـٌـ in the accusative, genitive and accusative respectively), typically used only in relatively formal language. For example, the tri-consonantal root /sˤdq/ is unambiguously read as /sˤidqun/ <صِدْقٌ> when the appropriate diacritics are added, with the first (rightmost) diacritic indicating a short /i/ vowel, the second indicating the absence of an immediately following short vowel, and the third indicating the nominative case. The orthographic system in Arabic thus has a number of interesting features. In this study, the focus is on the processing consequences of the presence versus absence of short vowel diacritics in written language. Written forms with such diacritics are often referred to as "voweled," and the inclusion of such diacritics as "voweling."

Because of the situation just discussed, Arabic is one of the few languages which can be considered to have both deep (opaque) and shallow (transparent) orthography (Frost, Katz, & Bentin, 1987; Katz & Frost, 1992). In the case of Arabic, deep orthography corresponds to the absence of diacritics and shallow orthography to their presence. More generally, deep orthography has been described as the "inconsistency between grapheme to phoneme, or the underspecification of phonological features" (Miller, Kargin, & Guldenoglu, 2014). For example, the unvoweled Arabic word-form <كتب> /ktb/ has a number of possible "readings" (i.e. pronunciations), including /kataba/ <كَتَبَ> "he wrote" and /kutubun/ <كُتُبٌ> "books." From these, the reader must choose an appropriate one based on context, consistent with a deep orthographic system. However, it is also true that each grapheme-to-phoneme mapping is consistent with respect to the graphemes actually present, with the associated phonological features all well-specified, more typical of shallow orthography. Furthermore, some forms have only one possible reading; an example of this is <سماء> "sky", which can only be read as /samaaʔ/ when case is not marked. These facts might therefore lead one to question whether Arabic genuinely has both deep and shallow orthography.

These issues are complicated further because of the situation of diglossia prevailing in the Arabic-speaking world. Saiegh-Haddad (2005) notes that written letters may sometimes map onto different phonemes in Modern Standard Arabic (MSA) and a given reader's native dialect. In some cases, the MSA phoneme might not even be a part of the reader's native phonemic inventory. This means that even

*Department of English Literature and Linguistics, College of Arts and Sciences, Qatar University, PO Box 2713, Doha, Qatar. For questions, contact Michael Grosvald at mgrosvald@qu.edu.qa. The authors would like to thank Leen Alsayed Hasso for assistance with stimuli creation and data collection. This research was supported by a Qatar University grant awarded to the first author (QUSG-CAS-DELL-14\15-24).

 Proceedings of the 36th West Coast Conference on Formal Linguistics, ed. Richard Stockwell et al., 176-181. Somerville, MA: Cascadilla Proceedings Project.

fully voweled forms might not really be shallow from the point of view of that individual. Taking all of these factors into account, we find it reasonable to suggest that at least in certain contexts, the Arabic writing system exhibits characteristics of both deep and shallow orthography.

The example of the written form <كتب> /ktb/, given earlier, illustrates the concept of "ambiguity." In the present context, a form is ambiguous if it has multiple readings when written without diacritics, while non-ambiguous forms necessarily have only one reading, whether written with diacritics or without. If ambiguous words are written with diacritics, immediate recognition is expected to occur on the part of the reader. However, even in the absence of diacritics, the effects of ambiguity can be mitigated by factors such as context and the relative frequencies of the readings of the word.

Bentin and Frost (1987) examined lexical decision and word naming performance in Hebrew, whose orthography is similar to Arabic's in having deep (unvoweled) and shallow (voweled) alternatives. Word-forms were presented in isolation. Lexical decisions were faster for unvoweled ambiguous words than for the corresponding voweled words. This was the case for both high-frequency and low-frequency voweled forms. In the word naming task, responses for ambiguous words were fastest for the unvoweled forms and the voweled high-frequency alternatives, and slowest for the voweled low-frequency alternatives. Collectively, these results suggest that the presence of vowel diacritics tends to make reading slower, but that readers do not (and perhaps cannot) ignore diacritics completely.

In a study of Arabic readers of various skill levels, Abu-Rabia (1997) examined how vowel diacritics and context affect reading accuracy. Two groups – 44 skilled readers and 34 poor readers – read words presented in isolation and in sentence and paragraph contexts. Word stimuli were shown in three different text conditions: unvowelized, partially vowelized, and fully vowelized. The results showed that (a) the presence of vowel diacritics facilitated reading for both groups in all conditions, and that (b) this benefit was greatly reduced in the sentence and paragraph conditions. These findings indicate that context plays an important disambiguating role for unvoweled word-forms.

Abu-Rabia (1998) presents the results of a similar study, again focusing on native Arabic speakers classified as either skilled or unskilled readers. Different kinds of Arabic written texts were presented; these included narrative, informative, poetic and Koranic. Each kind of text was presented in three orthographic conditions: unvowelized or with correct or incorrect diacritics. Both groups of readers showed better performance on the correctly vowelized Arabic forms than on the unvoweled texts, and incorrectly voweled texts led to the lowest scores of all.

Abu-Rabia's (2001) study on voweling in Arabic and Hebrew found that recognition of vowelized words in Arabic was more accurate than that of unvowelized words, regardless of whether the words were presented in isolation or in a full-paragraph context. In the unvowelized condition, accuracy was better in the full-paragraph context than when words were presented in isolation, again showing that context plays an important role in disambiguating unvoweled Arabic words. In contrast, accuracy for vowelized words did not significantly differ between the full-paragraph and isolated words contexts, due to the "shallow orthography effect."

The results of imaging research also support the notion that when vowel diacritics are included in written forms, readers incur some processing cost. An fMRI study by Bourisly and colleagues (2013) used a lexical decision task to investigate the neural correlates of the processing of diacritics by adult Arabic readers. There was no significant effect of voweling with respect to reading accuracy, but there was a significant effect on reaction times; unvoweled words were recognized faster than voweled words. Unvoweled words were associated with activation in the hippocampal and middle temporal gyrus regions, perhaps indicative of the need to search among multiple word candidates. In contrast, voweled words elicited activation in the insula and inferior frontal regions, which the researchers said might reflect phonological and semantic processes involved in recognizing voweled words. The researchers claimed these results "support a linguistic role for diacritics in isolated word recognition even in experienced Arabic readers."

Finally, Hermena and colleagues (2015) carried out an eye-tracking experiment using ambiguous Arabic verb forms that could allow either an active- or passive-voice reading. Verbs were shown within a sentence context, either with or without vowel diacritics. When only the ambiguous form was voweled, and the rest of the sentence left unvoweled, diacritics had a disambiguating effect that benefited readers. In the complete absence of vowel symbols, the default (typically, active voice) reading was chosen,

leading to garden-path effects when the appropriate reading was passive. Finally, when entire sentences were fully voweled, the result was taxing and readers did not seem to fully process the diacritics, reducing the benefit from the presence of diacritics on passive forms.

The present research study was designed to answer two research questions. The first of these was: will the presence of diacritics on word-forms presented in isolation affect readers' ability to perform a lexical decision task? Our expectation is that diacritics should delay responses in such a task, in harmony with the findings of a number of previous studies, as discussed earlier.

We also expect that the presence of diacritics will tend to reduce accuracy for responses to pseudo-words. This is because a pseudo-word may tend to look like a word superficially when written with diacritics, with the reverse not as likely. This first impression may be particularly likely to lead to errors when a reader must view such a form and make a quick decision about its lexical status.

Our second research question was: will readers' ability to perform a lexical decision task be influenced by whether a word-form is ambiguous or not? On the one hand, ambiguous forms may enable quicker lexical access because more than one lexical entry is available for such forms, leading to faster RTs. On the other hand, non-ambiguous word-forms might be expected to lead to faster RTs, in that they have only one reading in the lexicon. This might mean that they require less time to process, in turn helping participants to recognize such forms faster.

2. Methods

2.1. Participants

Forty-one female students at Qatar University took part. All reported normal or corrected-to-normal vision, no cognitive/linguistic impairments, and good knowledge of Modern Standard Arabic. Subjects' native dialects were as follows: Kuwaiti (1), Syrian (2), Sudanese (4), Bahraini (5), Saudi (7), Omani (3), Yemeni (2), Egyptian (2), Palestinian (4), Qatari (4), Jordanian (2), Moroccan (1), and Tunisian (2). Two subjects reported being bilingual in Arabic and one other language (English, Somali). Study participants received no compensation for taking part in the study, and provided informed consent in accordance with established Institutional Review Board procedures at Qatar University.

2.2. Materials

In order to create a set of suitable stimuli, Modern Standard Arabic dictionaries were first used to collect a group of several dozen words with tri-consonantal roots. Next, these words were categorized as ambiguous or non-ambiguous by a group of native Arabic speakers, as described in the "Pre-test" section just below. These individuals were different from those who took part in the main experiment. Finally, the web resource Aralex (Boudelaa & Marslen-Wilson, 2010) was used to ensure that word frequencies were balanced between the ambiguous and non-ambiguous conditions. The final set of 80 stimuli consisted of 40 real words, evenly balanced between ambiguous and non-ambiguous (see below), as well as 40 phonologically legal pseudo-words.

2.3. Pre-test

Eighteen native Arabic speakers took part in this pre-test, whose goals were to determine (1) whether a given word could legitimately be classified as ambiguous or not, and (2) for ambiguous words, what the most common reading was. During this pre-test, the candidate word stimuli were presented on-screen one-by-one in random order, in unvoweled form. For each such stimulus, the speaker was asked to say aloud all of the possible readings that came to mind. For example, for unvoweled /ktb/ <كتب>, the possible readings include /kataba/ <كَتَبَ> "he wrote" and /kutubun/ <كُتُبٌ> "books", among others. These sessions were audio-recorded and transcribed.

In isolation, Arabic verbs and nouns can be pronounced with or without a final vowel, (/u(n)/, ـٌ) for nouns and (/a/, ـَ) for verbs, and some pre-test participants produced such endings while others did not. However, for our purposes these variations were considered unimportant, as the distinguishing factor in determining ambiguity was the (inter-consonantal) vowel pattern.

2.4. Main experiment

The data were collected in a quiet room on the Qatar University campus. Each participant was seated facing a laptop screen, and was given a description of the experimental task and a basic set of instructions. Before starting the actual experiment, each participant completed a warm-up version of the real experiment and was encouraged to ask for clarifications if any were needed.

The experiment was programmed using Presentation software (Albany, CA, US), which presented the series of stimuli to participants and logged their responses. Participants viewed the series of 80 word and pseudo-word stimuli on-screen one by one in a randomized order. For each stimulus item, the participant needed to decide whether the item was a word or a pseudo-word, indicating their decision by pressing the keyboard key "Z" for "word" and "/" for "pseudo-word." Each word and pseudo-word could be presented with or without diacritics, as determined at random for each item trial by trial. For ambiguous words presented with diacritics, the form used was the most common reading as determined by the Pre-test results.

2.5. Data analysis

The dependent measures we examined are accuracy and response times (RTs) for word and pseudo-word stimuli. For both measures, we conducted repeated-measures ANOVAs by subjects (F1) and items (F2), performing further (planned and post-hoc) statistical comparisons when relevant using t-tests. In the RT analyses, we used only data from trials for which the answer was correct. Accuracy scores were analyzed after applying an arcsin transformation ((y=arcsin(sqrt(x)), but for clarity of presentation, accuracy results are given with the data back-transformed into percentage scores (Cook & Wheater, 2005). For significant outcomes of ANOVA analyses, we report both F-statistics and their associated p-values, along with mean square error (MSE) and effect size (Δ, i.e. "difference"). Because all of our factors of interest had two levels only, corrections to address issues of sphericity (e.g. Huynh-Feldt, Greenhouse-Geisser) were not needed.

The three factors of interest were Lexicality (word or pseudo-word), Diacritics (present or absent), and Ambiguity (ambiguous or non-ambiguous), the last of which could only apply to real words, not pseudo-words. Accordingly, for each of the two dependent measures, the following pair of 2-by-2 ANOVAs were performed. The first was on the factors of Lexicality and Diacritics, and the second was on the factors of Diacritics and Ambiguity, with this second ANOVA restricted to words only.

3. Results

The accuracy ANOVA with factors Lexicality and Diacritics yielded a highly significant main effect of Lexicality ($F1(1,40)=40.246$, $MSE=0.015$, $p<0.001$; $F2(1,77)=19.341$, $MSE=0.033$, $p<0.001$) and a main effect of Diacritics that was significant only by items ($F1<1.5$; $F2(1,77)=7.602$, $MSE=0.012$, $p<0.01$). There was no significant interaction ($Fs<1$).

The effect of Lexicality is due to the fact that responses to words were overall more accurate than those to pseudo-words (99.5% vs 95.7%, Δ=3.8%). The effect of Diacritics reflects the fact that responses to stimuli with diacritics were somewhat *less* accurate overall than responses to stimuli which did not incorporate diacritics (97.7% vs 98.3%, Δ=0.6%). Although this is a marginal effect (significant only by items, as noted above), it foreshadows a finding that shows up more robustly in the reaction time data – namely, that the presence of diacritics appears to *inhibit* rather than *facilitate* subjects' performance on this word-reading task.

The accuracy ANOVA restricted to words only, with factors Ambiguity and Diacritics, found an effect of Ambiguity that was significant by subjects but not items ($F1(1,40)=5.177$, $MSE=0.018$, $p<0.05$; $F2<1.5$), and an interaction that reached significance by subjects but only marginal significance by items ($F1(1,40)=6.778$, $MSE=0.018$, $p<0.05$; $F2(1,37)=3.976$, $MSE=0.009$, $p=0.054$). The effect of Diacritics did not reach significance ($F1<1$; $F2(1,37)=2.813$, $MSE=0.009$, $p=0.102$).

The effect of Ambiguity is due to the fact that responses to ambiguous words were more accurate than those to non-ambiguous words (99.8% vs 99.1%, Δ=0.7%). The interaction reflects the fact that the

influence of diacritics was different for ambiguous words, where the presence of diacritics was associated with significantly reduced accuracy (99.9% vs 99.4%, Δ=0.6%; $p<0.05$), compared to non-ambiguous words, for which this difference was not significant, but showed a trend in the opposite direction (98.6% vs 99.5%, Δ=0.9%; $p=0.17$).

The RT ANOVA with factors Lexicality and Diacritics found highly significant main effects of both Lexicality ($F1(1,40)=48.746$, MSE=22428, $p<0.001$; $F2(1,77)=105.629$, MSE=32406, $p<0.001$) and Diacritics ($F1(1,40)=34.246$, MSE=22428, $p<0.001$; $F2(1,77)=60.025$, MSE=22966, $p<0.001$), as well as an interaction that reached significance by subjects but showed only a marginal trend by items ($F1(1,40)=4.306$, MSE=22428, $p<0.05$; $F2(1,77)=2.341$, MSE=22966, $p=0.13$).

The effect of Lexicality is due to the fact that responses to words were quicker than responses to pseudo-words (893 vs 1202 ms, Δ=309 ms). The effect of Diacritics shows that responses to stimuli were faster when diacritics were absent versus when they were present (945 vs 1149 ms, Δ=204 ms). The presence of an interaction reflects the fact that this slowing of responses in the presence of diacritics was of greater magnitude for pseudo-words (1076 vs 1328 ms, Δ=252 ms, $p<0.001$) than for words (815 vs 970 ms, Δ=155 ms, $p<0.001$).

Finally, the RT ANOVA for words only, with factors Ambiguity and Diacritics, yielded a highly significant effect of Diacritics ($F1(1,40)=22.776$, MSE=13890, $p<0.001$; $F2(1,37)=32.622$, MSE=13348, $p<0.001$). There was no significant effect of Ambiguity ($Fs<1$) and no interaction ($Fs<2$). The effect of Diacritics is due to the fact that responses to words were slower when diacritics were present, as discussed just above.

4. Discussion

The main purpose of the present study is to investigate the influences of vowel diacritics and orthographic ambiguity on Arabic readers' performance on a lexical decision task. Our data yield three main findings. First, consistent with earlier psycholinguistic studies, response times were faster to words than to pseudo-words. Second, although voweled forms provide more information than unvoweled forms, which might therefore be expected to facilitate lexical access, the presence of diacritics slowed our participants' response times greatly. Third, this slowing of response times in the presence of diacritics was observed for all conditions of lexicality and orthographic ambiguity.

The slowing associated with diacritics is consistent with findings from previous studies, which have often found that diacritics come with a processing cost. This is especially so for isolated word-forms, though diacritics can also facilitate the reading of ambiguous words in a sentence context. This suggests that the presence of short vowel symbols leads to increased visual crowding; everyday Arabic text contains either no or very few vowel diacritics. Abdelhadi, Ibrahim, and Eviatar (2011) argue that the Arabic orthographic system can itself be considered visually complex in general -- that is, independent of the presence of diacritics. The reasons they give include the multiple forms individual letters can take, the ways some letters connect to others, and the role that dots play in forming Arabic letters. From this standpoint, the presence of vowel diacritics in Arabic text can be seen as adding an additional burden to what is already a complex and challenging type of visual-linguistic processing.

This is consistent with the findings of the study discussed in Abdelhadi, Ibrahim, and Eviatar (2011), which examined the influence of orthographic visual complexity in the context of a detection task. Participants were native Arabic speakers in the third and sixth grades who were learning Hebrew as an L2 in Hebrew schools. Their task was to detect vowel diacritics in Arabic and Hebrew word-forms. Stimuli consisted of real words, orthographically legal pseudo-words, and "words" formed with non-linguistic "letters" which superficially resembled real letters. Participants actually performed better in the L2 task (i.e., Hebrew) than in the L1 task (Arabic), which the researchers attributed to the greater "orthographic visual complexity" of Arabic.

Eviatar and Ibrahim (2014) found a reduced contribution on the part of the right hemisphere in letter and word recognition in Arabic, relative to Hebrew and English. This may at least partially explain the finding that reading acquisition in young learners of Arabic is relatively slow. Even in adult readers, the researchers find, the right hemisphere "does not contribute to the early stages of reading, such as word recognition, as we know it does in other languages." The researchers argue that these findings can be

attributed to the complexity of Arabic orthography, as well as the situation of diglossia noted earlier. The issue of visual complexity also appears relevant to the findings of Mountaj and colleagues (2015), who conducted an experiment combining semantic priming and EEG recording. The presence of diacritics in written Arabic forms had no significant effect on semantic priming, but the N1 and N2 ERP components were larger when vowel diacritics were present, pointing towards an increased visual crowding effect.

A related finding in the present study was that the presence of diacritics showed differential influence on the accuracy of responses to words and pseudo-words, such that that participants found it harder to reject pseudo-words in the presence of diacritics. Some participants, when faced with a voweled form, may have erred on the assumption that this might be an existing MSA word with which they were unfamiliar. In addition, because vowel diacritics are associated with phonological and semantic information in the language, it is likely that their presence automatically triggers phonological and semantic processing.

The other main finding in the present study was the overall lack of strong effects and interactions related to word ambiguity. Accuracy of participants' responses were generally higher for ambiguous words, particularly those written without diacritics. In contrast, ambiguity appeared to have essentially no influence on the RT results. Instead, the dominating factor in those results was the presence (versus absence) of diacritics, which slowed response times to word stimuli to a consistent degree regardless of whether or not the words were ambiguous. We found no facilitatory effect of the presence of vowel diacritics on the accessing of ambiguous words presented in isolation. This suggests that in cases of ambiguity, readers simply fall back on the default reading, without paying attention to the diacritics.

References

Abdelhadi, Souad, Raphiq Ibrahim, & Zohar Eviatar (2011). Perceptual load in the reading of Arabic: Effects of orthographic visual complexity on detection. *Writing Systems Research*, *3*(2), 117-127.

Abu-Rabia, Salim (1997). Reading in Arabic orthography: The effect of vowels and context on reading accuracy of poor and skilled native Arabic readers. *Reading and Writing*, *9*(1), 65-78.

Abu-Rabia, Salim (1998). Reading Arabic texts: Effects of text type, reader type and vowelization. *Reading and Writing*, *10*(2), 105-119.

Abu-Rabia, Salim (2001). The role of vowels in reading Semitic scripts: Data from Arabic and Hebrew. *Reading and Writing*, *14*(1-2), 39-59.

Bentin, Shlomo, & Ram Frost (1987). Processing lexical ambiguity and visual word recognition in a deep orthography. *Memory & Cognition*, *15*(1), 13-23.

Boudelaa, Sami, & William D. Marslen-Wilson (2010). Aralex: a lexical database for Modern Standard Arabic. *Behavior Research Methods*, *42*(2), 481-487.

Bourisly, Ali K., Charles Haynes, Nibal Bourisly, & Maria Mody (2013). Neural correlates of diacritics in Arabic: An fMRI study. *Journal of Neurolinguistics*, *26*(1), 195-206.

Cook, Penny A., & P. Wheater (2005). *Using statistics to understand the environment*. London: Routledge.

Eviatar, Zohar, & Raphiq Ibrahim (2014). Why is it hard to read Arabic? In Elinor Saiegh-Haddad & R. Malatesha Joshi (Eds.), *Handbook of Arabic Literacy* (pp. 77–96). Springer Netherlands.

Frost, Ram, Leonard Katz, & Shlomo Bentin (1987). Strategies for visual word recognition and orthographical depth: a multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(1), 104.

Hermena, Ehab W., Denis Drieghe, Sam Hellmuth, & Simon P. Liversedge (2015). 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), 494.

Katz, Leonard, & Ram Frost (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. In Ram Frost & Leonard Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 67-84). Amsterdam: Elsevier Science Publishers.

Mountaj, Nadia, Radouane El Yagoubi, Majid Himmi, Faouzi Lakhdar Ghazal, Mireille Besson, & Sami Boudelaa (2015). Vowelling and semantic priming effects in Arabic. *International Journal of Psychophysiology*, *95*(1), 46-55.

Saiegh-Haddad, Elinor (2005). Correlates of reading fluency in Arabic: Diglossic and orthographic factors. *Reading and Writing*, *18*(6), 559-582.

Proceedings of the 36th West Coast Conference on Formal Linguistics

edited by Richard Stockwell, Maura O'Leary, Zhongshi Xu, and Z.L. Zhou

Cascadilla Proceedings Project Somerville, MA 2019

ISBN 978-1-57473-474-4 hardback

Ordering information

Orders for the printed edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Grosvald, Michael, Sarah Al-Alami, and Ali Idrissi. 2019. Word Reading in Arabic: Influences of Diacritics and Ambiguity. In *Proceedings of the 36th West Coast Conference on Formal Linguistics*, ed. Richard Stockwell et al., 176-181. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #3460.