Orthographic ambiguity, diacritic density, and morphology in Arabic word reading

Ali Idrissi[1], Eiman Mustafawi[1],

Tariq Khwaileh[1] & R. Muralikrishnan[2]

[1] *Qatar University*

[2] *Max Planck Institute for Empirical Aesthetics*

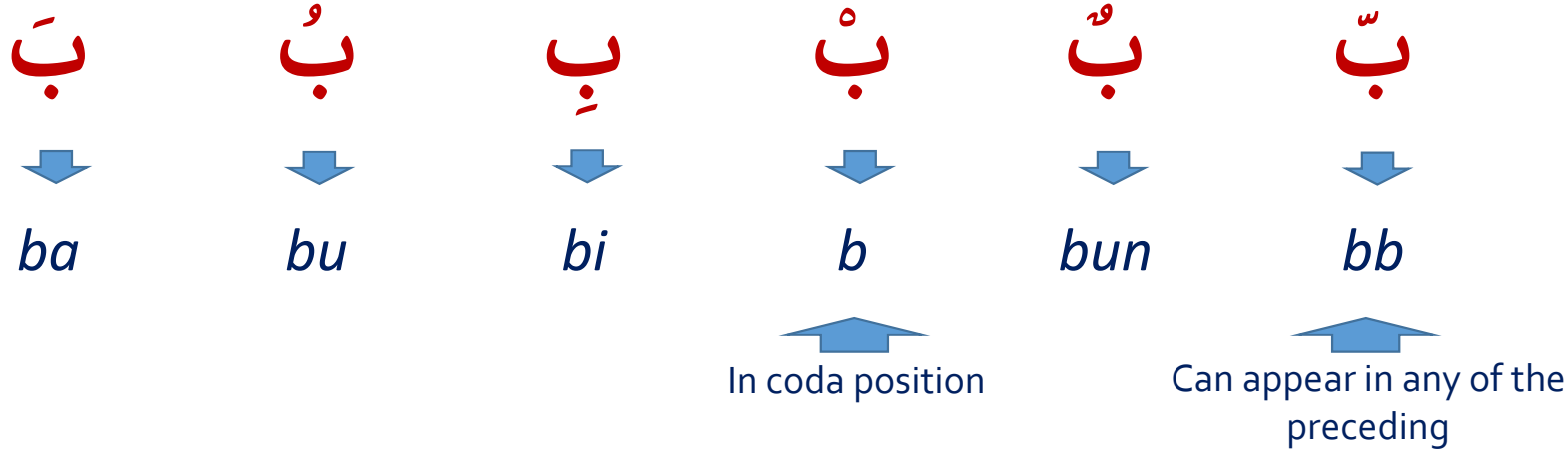
Research supported by QNRF grant # NPRP 7-427-6-011

Outline

- Arabic script
- Some previous work
- The current study
- Methods
- Results
- Discussion and conclusions

Arabic Script

- Independent symbols only for **consonants** and **long vowels**.
- Short vowels, gemination, case endings, and 'absence of vowel' are indicated with **diacritical** marks above and underneath consonant symbols:



Orthographic Depth

(Frost, Katz, & Bentin, 1987; Katz & Frost, 1992)

- In everyday text, diacritics are left out.
- Thus, Arabic orthography show two forms:

Deep (opaque)

No straightforward grapheme-phoneme correspondence.

Shallow (transparent)

Some form of grapheme-phoneme correspondence, due to diacritics.

With the deep form being more **natural** for skilled readers.

Orthographic Depth

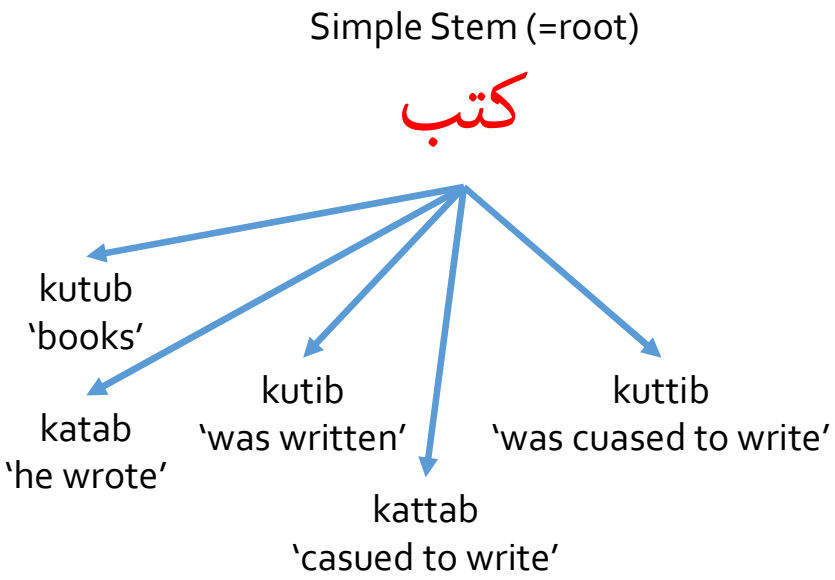
- Skilled readers exposed to 'nondiacriticized' text, with no indication of:
  - short vowels (or of absence of a vowel)
  - gemination
  - or case

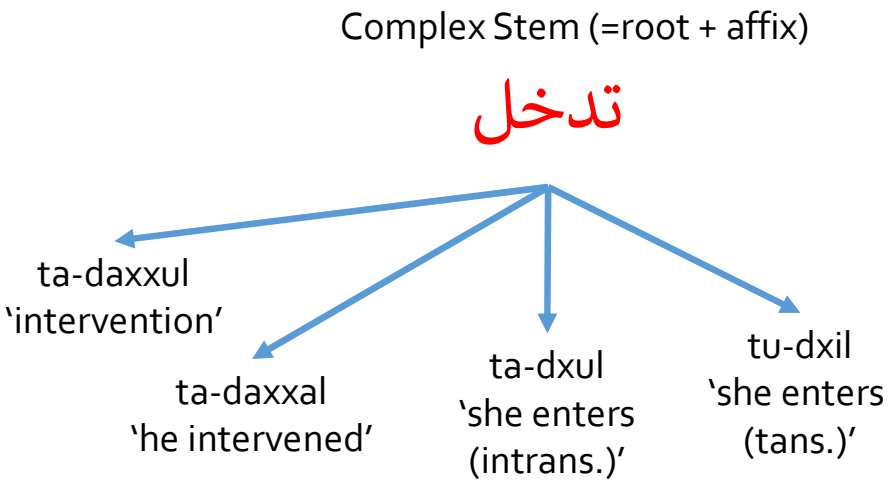
الإندبندنت: مرسي ترك ملقى على الأرض لأكثر من 20 دقيقة

قالت صحيفة الإندبندنت البريطانية إن الشرطة المصرية متهمة بالتسبب في قتل الرئيس المصري المعزول محمد مرسي، بعدما تركته ملقى على أرضية قفص الاتهام في المحكمة لمدة عشرين دقيقة.

Orthographic Un/Ambiguity

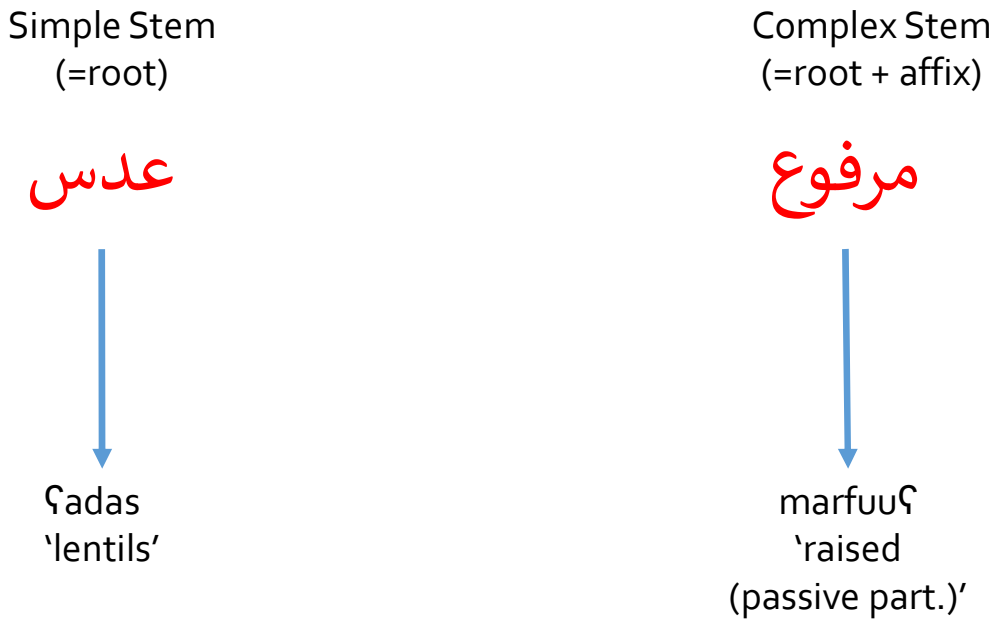
- Lack of diacritics leads to a prevalence of **heterophonic homographs**:
- Written forms of simple or complex stems with more than one possible pronunciation/meaning:





Orthographic Un/Ambiguity

- Along with unambiguous written forms (with one possible pronunciation/meaning):



Degrees of diacriticization

No diacriticization (NON)

NON-diacriticized mode

- The most natural way to present text in Arabic.

الإندبندنت: مرسي ترك ملقى على الأرض لأكثر من 20 دقيقة

قالت صحيفة الإندبندنت البريطانية إن الشرطة المصرية متهمة بالتسبب في قتل الرئيس المصري المعزول محمد مرسي، بعدما تركته ملقى على أرضية قفص الاتهام في المحكمة لمدة عشرين دقيقة.

Full diacriticization (FULL)

- The **least natural** way to present text in Arabic, typical of religious, educational and some literary texts

Hadith "Prophet's saying"

School textbook

FULLy-diacriticized mode

Optimal diacriticization (MIN)

- Probably, because syntactic/pragmatic contexts cannot always help, minimal diacriticization is often used for disambiguation purposes:

عزّز وزير الخارجية البريطاني السابق بوريس جونسون تقدّمه في السباق لخلافة رئيسة الوزراء البريطانية تيريزا ماي، بعدما تصدّر نتائج الدورة الثانية من تصويت النواب المحافظين.

وستجرى جولات أخرى لتقليص القائمة إلى أن يتبقى اثنان فقط يُنتخب أحدهما من طرف كل منتسبي الحزب في أنحاء البلاد، البالغ عددهم 160 ألفًا.

Aljazeera.net news page

يا عَبلَ حبُّكِ في عِظامي مَع دَمي لَمّا جرَت روحي بجِسمي قَد جَرى

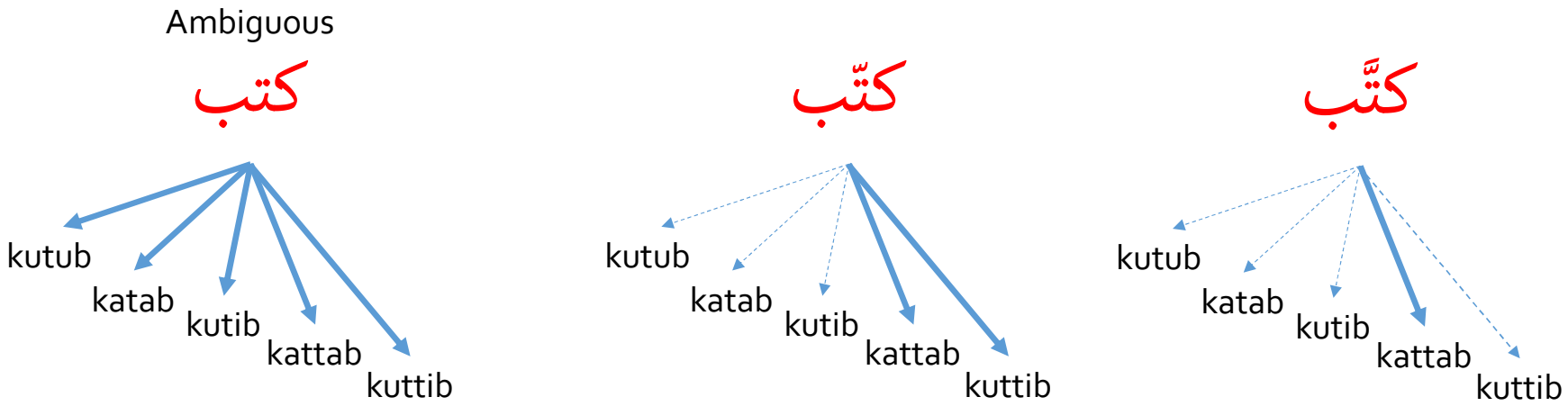
Line from Antar Bnu Shaddad's poem

Minimal diacriticization seems to be aimed at reducing ambiguity as much as possible.

MINimally- diacriticized mode

Diacritic density

- Disambiguation can be achieved through minimal/optimal diacriticization:



Diacritic density

- To our knowledge there are no studies dealing with the relationship between diacritic density and either reading accuracy or reading times.
- We selected three levels:
  - Full (fully diacriticized) **F**
  - Optimal (minimally diacriticized) **O**
  - Non (non-diacriticized) **N**

Questions

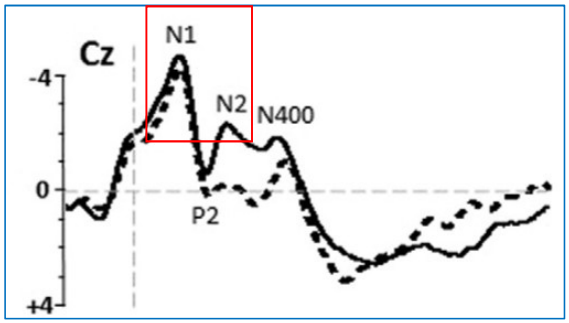
- In such a system with at least three degrees of depth/shallowness:

1) How do readers **perform** the task of retrieving the correct reading of isolated words and words in context?
2) Because they represent important (otherwise missing) information, would diacritics **increase accuracy** during this task?
3) Because they are typically absent (i.e., not frequent), would diacritics affect the **speed** with which the task is performed?
4) Does orthographic **ambiguity** matter?
5) Does the **amount** of diacriticization matter?

Previous findings: **Reading speed**

- Reports converge on a ***processing cost*** to diacritics:
- A slowing effect of diacritics on reading/recognition speed:
  - Arabic (Bourisly et al. 2013; Hermena et al. 2015; Grosvald & Idrissi *in review*).
  - Hebrew: Bentin & Frost (1987) showed the same for pointed vs. unpointed forms.

- Visual noise:
  - ERPs: larger N1 and N2 in diacriticized words) (Mountaj et al. 2015)



Previous findings: **Reading accuracy**

- ***Conflicting results**:*
- Diacritics **improve** reading accuracy of isolated words (Abu-Rabia 2001 for a review)
- Diacritics **reduce** reading accuracy of isolated words (Abu-Leil et al. 2014; Idrissi & Grosvald *in review*).

Orthographic Ambiguity

- But, do diacritics have the same effect on AMB and UMB words?
- Maroun & Hanley (2017):
  - diacritics **increased** accuracy on ambiguous words compared to unambiguous words in isolation and in a sentence context
- In a priming study, Idrissi & Grosvald (in preparation) found that:
  - diacritics **decreased** accuracy on ambiguous words.

Conflicting results, but different tasks!

Orthographic Ambiguity

- Hermena et al. (2015): Eye-tracking reading ambiguous Arabic verb forms (active or passive-voice)
- In a sentence context, with (FULL) or without (NON) vowel diacritics.
- Results:
  - When only AMB form was voweled, disambiguating happened.
  - When it was not, the verb was read in the default (active) voice (> garden-path effects in the passive context).
  - When the whole sentence was voweled, diacritics were not only taxing but readers seemed to ignore them.

Current picture

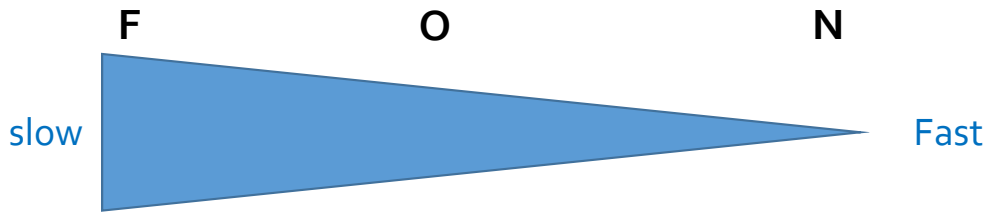
- Syntactic context does not seem to help; readers fall back on the default reading (see eye tracking results).
- Diacritics are taxing and require more processing/attention.
- Conflicting results on accuracy (beneficial and disruptive for **word** lexical retrieval/reading)
- Possibly conflicting results on the accuracy and diacritics in the context of ambiguity (beneficial and disruptive of **ambiguous word** lexical retrieval/reading)

Our questions

- Clear processing cost of diacritics,
- Their exact impact is on **reading accuracy** in the context of orthographic **ambiguity** remains unclear.
- Questions:
  - Effect of **diacritics** on word **reading accuracy**?
  - Effect of **ambiguity** on **reading accuracy** and **speed**?
  - Effect of **diacritics** on **reading accuracy of ambiguous words**?
  - Relationship between **diacritic density** and reading accuracy and speed?
  - Effect of **frequency** on reading accuracy and speed?
  - Also, since the more complex a stem, the less ambiguous it is, would **stem complexity** matter?

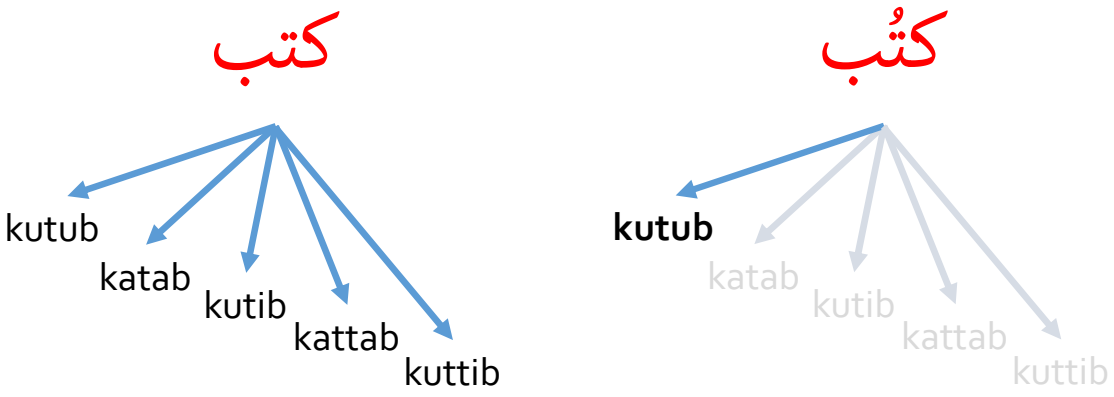
Hypotheses: Reading times

- Reading Times (RTs) and diacritic density:
- Since diacritics may be mere visual noise and tend to slower word reading speed, we should see a parametric effect of diacritic density on reading times:



Reading accuracy: Hypothesis 1

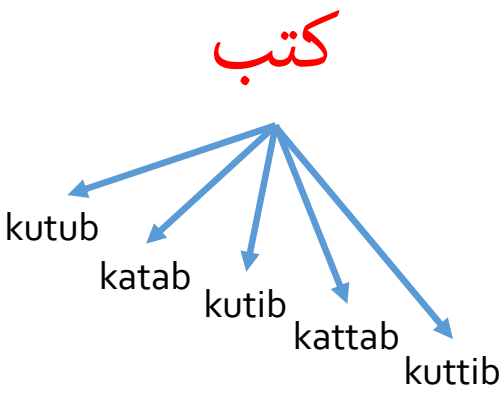
- Since diacritics seem to facilitate access to semantic representations,
- and since an ambiguous form activates more than one candidate reading,
- Diacritics should be more beneficial in the case of ambiguous than unambiguous words.

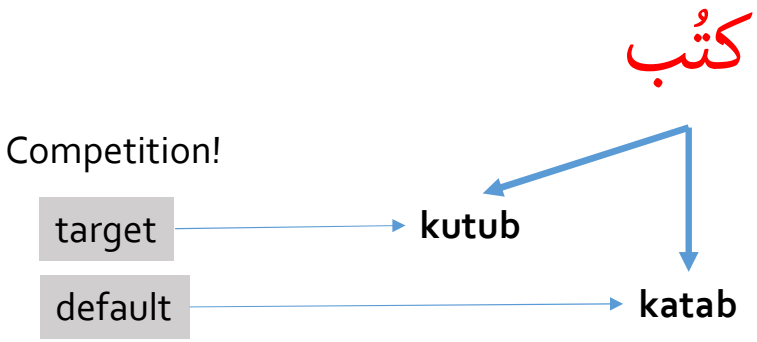


Consistent with Abu-Rabia (2001), Abu-Leil et al. 2014), and is not inconsistent with Maroun & Hanley (2017)

Reading accuracy: Hypothesis 2

- Given Arabic morphology and its lexical organization, diacritics (regardless of their density) should 'disrupt' the normal AMB word reading processes (automatic combination of the root and the **default** associated word pattern), leading to errors at:
  - the level of morphological composition
  - or the grapheme-to-phoneme conversion stage





Hypotheses

- Stem complexity:
  - Since reading is assumed to be guided by the root, we do not expect any effects of stem complexity.

- Stem frequency:
  - Since reading ambiguous words is assumed to be further guided by the default vs. non-default reading, we do not expect any effects to stem frequency.

Methods: Materials

- 144 words:
  - 72 ambiguous
  - 72 non-ambiguous
  - 50% high frequency
  - 50% low frequency
  - 50% simple stem
  - 50% complex
- All appeared as:
  - Fully voweled (FULL)
  - Optimally/partially voweled (MIN)
  - Zero-voweled (NON).
- Total: 432 word forms.

Methods: conditions

	Ambiguous			***Unambiguous***		
	N	***O***	***F***	***N***	***O***	***F***
Simple stems:	كتب	كتب	كَتَبَ	عدس	عدَس	عَدَسٌ
Complex stems:	تدخل	تدخُّل	تَدَخُّلٌ	مرفوع	مَرفوع	مَرْفُوعٌ

Ambiguous, simple and complex

Methods: conditions

	Ambiguous			*Unambiguous*		
	N	***O***	***F***	***N***	***O***	***F***
Simple stems:	كتب	كتَب	كَتَبَ	عدس	عدَس	عَدَسٌ
Complex stems:	تدخل	تدخُّل	تَدَخُّلٌ	مرفوع	مَرفوع	مَرْفُوعٌ

Unambiguous: simple and complex

Methods: conditions

	Ambiguous			*Unambiguous*		
	N	***O***	***F***	***N***	***O***	***F***
Simple stems:	كتب	كتَب	كَتَبَ	عدس	عدَس	عَدَسٌ
Complex stems:	تدخل	تدخُّل	تَدَخُّلٌ	مرفوع	مَرفوع	مَرْفُوعٌ

Non-diacriticized

Methods: conditions

	Ambiguous			*Unambiguous*		
	N	***O***	***F***	***N***	***O***	***F***
Simple stems:	كتب	كتَب	كَتَبَ	عدس	عدَس	عَدَسٌ
Complex stems:	تدخل	تدخُّل	تَدَخُّلٌ	مرفوع	مَرفوع	مَرْفُوعٌ

Fully-diacriticized

Methods: conditions

	Ambiguous			***Unambiguous***		
	N	***O***	***F***	***N***	***O***	***F***
Simple stems:	كتب	كتَب	كَتَبَ	عدس	عدَس	عَدَسٌ
Complex stems:	تدخل	تدخُّل	تَدَخُّلٌ	مرفوع	مَرفوع	مَرْفُوعٌ

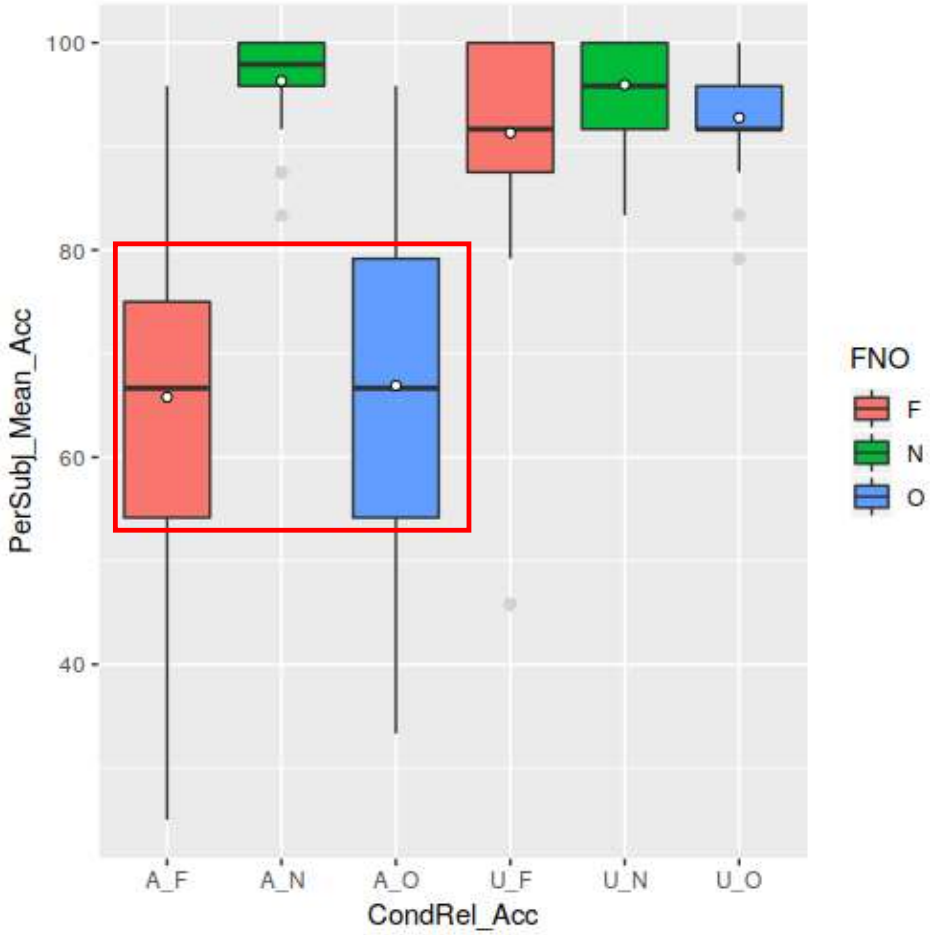
Minimally-diacriticized

Participants and procedure

- 34 adult, university-educated native speakers of Arabic
- Task: Read each word aloud, as accurately and as quickly as possible.
- Responses and response times were recorded.
- Statistical analysis
- Error analysis

Results: Reading Accuracy

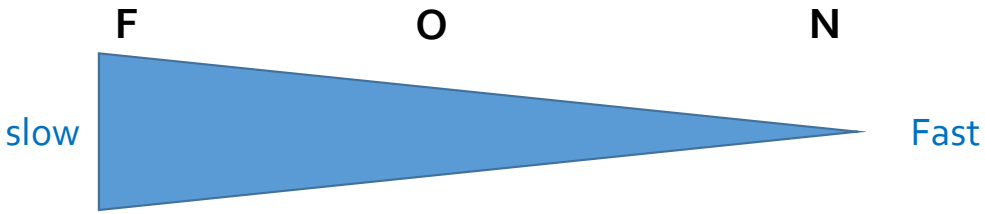
- Reading accuracy of UMB is higher and insensitive to DD.
- Reading accuracy of AMB words decreases in F and O; but was like UMB in N condition
- So, we have an interaction between ambiguity and diacritic density.
- In sum:
  - Diacritics **reduce** reading accuracy for ambiguous words.

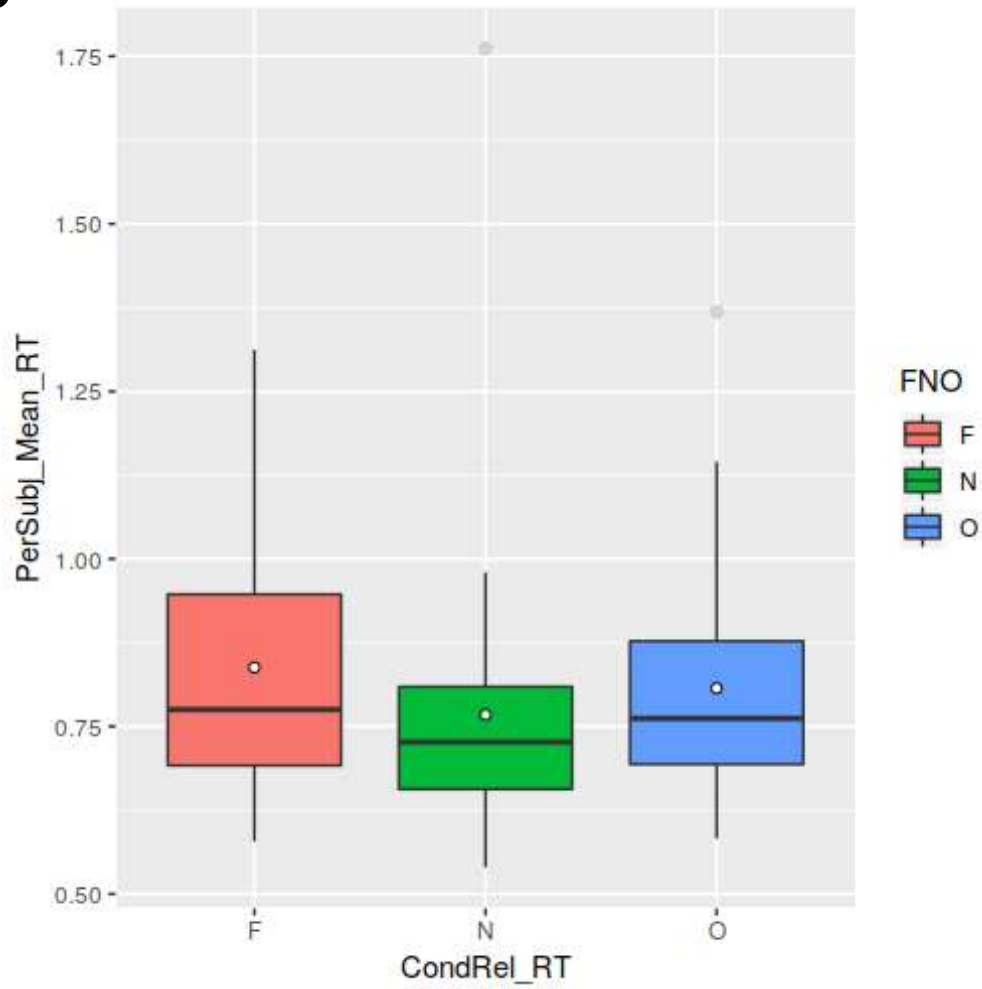


Results: Response Times

- Response times are always sensitive to diacritic density, regardless of ambiguity.

- Parametric effect:
  - F read less fast than O, which is read less fast than N.





Summary of results

- Reading **accuracy**:
  - Ambiguity interacts with diacritic density.
    - When diacriticized, ambiguous words read less accurately than diacriticized unambiguous words.
    - NON read most accurately, FULL and MIN read less so.
- Reading **times**:
  - Only diacritic density affected reading times.
  - RTs as a function of diacritic density: FULL slowest, NON fastest, MIN in between.

Discussion

- **Reading/recognition speed**:
  - We confirm the slowing/taxing effect of diacritics, but add more support for their visual 'noise' nature by showing the correlation between DD and RTs.

- **Reading accuracy**:
  - Diacritics **decrease** reading accuracy for AMB words ONLY.
  - And they do **not seem to matter** in the case of UMB words (all read accurately regardless of diacriticization).

Discussion: Why?

- **Morphology and default reading precede phonology:**
- In Arabic, reading (and word recognition) is guided by morphology
  - (see data from both brain and behavioral experiments).
- When read, AMB words take the default reading
  - (see Hermana et al. 2015 for passive > active).
- Phonology (or grapheme to phoneme conversion) is **not** the natural route for skilled readers
  - (Simon et al. 2006 and the absence of N320 in Arabic reading compared to French).

Discussion

- FULL and MIN suppresses the default reading (which involves root-WP combination).
- Reader is prompted to either:
  - ***Option 1:*** "look up" the target pronunciation/meaning among the possible candidates/the set of competitors, or
  - ***Option 2*** (rare): attempt grapheme-to-phoneme conversion

Discussion

- **These options predict:**
- **Option 1:** morphological (or morpho-orthographic) errors.

  e.g., رَقْمٌ [raqm] 'number' → [raqqam] 'to number/give a number to'

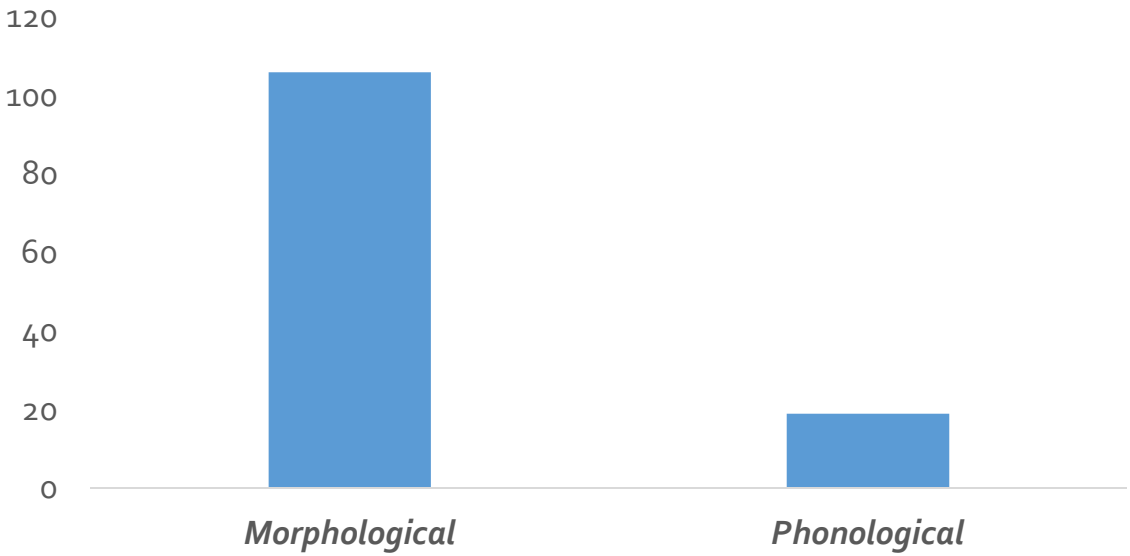
  مَرْكَبٌ [markab] 'boat' → [murakkab] 'complex'

- **Option 2:** occasional mispronunciation errors

  e.g., غَرْبٌ [ɣarbun] 'west' → * ɣarabun

Discussion

- Predictions are borne out by the data (type and proportion of errors)
- Error analysis shows that the majority of errors on AMB words were morphological (same root, but different word pattern):



Conclusions

- Diacritics do constitute some form of 'visual' noise for skilled readers.
  - In line with previous ERP and Eye-tracking evidence.
  - More evidence from the graded/parametric difference between F, O and N.
- Diacritics interact with ambiguity:
  - They do not matter in reading UMB words.
  - But, when present, they reduce accuracy in reading AMB words.
- Explanation:
  - We argue for the prevalence of the morpho-orthographic route in reading Arabic (Bar-On et al. 2018)
- Further research is yet to shed light on how much phonology is still involved in reading Arabic and at what stage it yields in to morphology.

Merci.

Research supported by the Qatar National Research Fund grant # NPRP 7-427-6-011