# The Role of Diacritics in Adapting the Difficulty of Arabic Lexical Recognition Tests

**Osama Hamed**
Language Technology Lab
University of Duisburg-Essen
osama.hamed@uni-due.de

**Torsten Zesch**
Language Technology Lab
University of Duisburg-Essen
torsten.zesch@uni-due.de

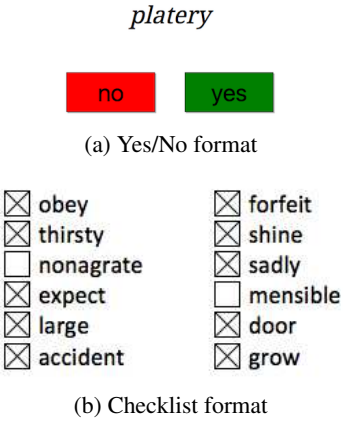
## Abstract

Lexical recognition tests are widely used to assess the vocabulary size of language learners. We investigate the role that diacritics play in adapting the difficulty of Arabic lexical recognition tests. For that purpose, we implement an NLP pipeline to reliably estimate the frequency of diacritized word forms. We then conduct a user study and compare Arabic lexical recognition tests in three settings: (i) without diacritics, (ii) with the most frequent diacritized form of a root, and (iii) the least frequent diacritized form of a root. We find that the use of infrequent diacritics can be used to adapt the difficulty of Arabic lexical recognition tests and to avoid ceiling effects.

## 1 Introduction

Lexical recognition tests (LRTs) are used to measure the vocabulary size of a learner. For that purpose, learners are presented with lexical items and have to decide whether they are part of the vocabulary of a given language (i.e. a *word*) or not (i.e. a *nonword*). Figure 1 gives an example of the two most common presentation formats: (i) Yes/No questions and (ii) checklists. A lexical recognition test consists of a relatively small number of words and nonwords, usually 40 words and 20 nonwords. It has been shown that such a small number of items is sufficient to consistently measure the vocabulary size (Huibregtse et al., 2002). As a consequence, lexical recognition tests are easy to administer and fast (Lemhöfer and Broersma, 2012).

Nonwords in a lexical recognition test are typically used as distractors. Thus, they should be close to existing words and are usually created by swapping letters in existing words (Stubbe, 2012) or by generating character sequences based on position-specific character language models (Hamed and Zesch, 2015). Words in a lexical recognition test have the function to measure the vocabulary size, thus the test needs to contain words from many frequency bands, i.e. very frequent words like *door* or *large* as well as less common words like *obey* or *forfeit*.




(a) Yes/No format

(b) Checklist format

Figure 1: Examples of lexical recognition tests.

While lexical recognition tests are well-established for English (Lemhöfer and Broersma, 2012), and other European languages like German and Dutch (Lemhöfer and Broersma, 2012), French (Brysbaert, 2013) and Spanish (Izura et al., 2014), there is still very little work on Arabic LRTs. The studies by Baharudin et al. (2014) and Ricks (2015) neglect lexical diacritics, a very important feature of the Arabic language that causes many challenges for automatic processing (Farghaly and Shaalan, 2009).

The Arabic script contains two classes of symbols: letters and diacritics (Habash, 2010). Whereas letters are always written, diacritics are optional. Diacritics are usually used in specific settings like language teaching or religious texts. This leads to a high amount of ambiguity of a non-diacritized Arabic word. Figure 2 compares the



Osama Hamed and Torsten Zesch 2018. The role of diacritics in increasing the difficulty of Arabic lexical recognition tests. *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning at SLTC 2018 (NLP4CALL 2018)*. Linköping Electronic Conference Proceedings 152: 23–31.

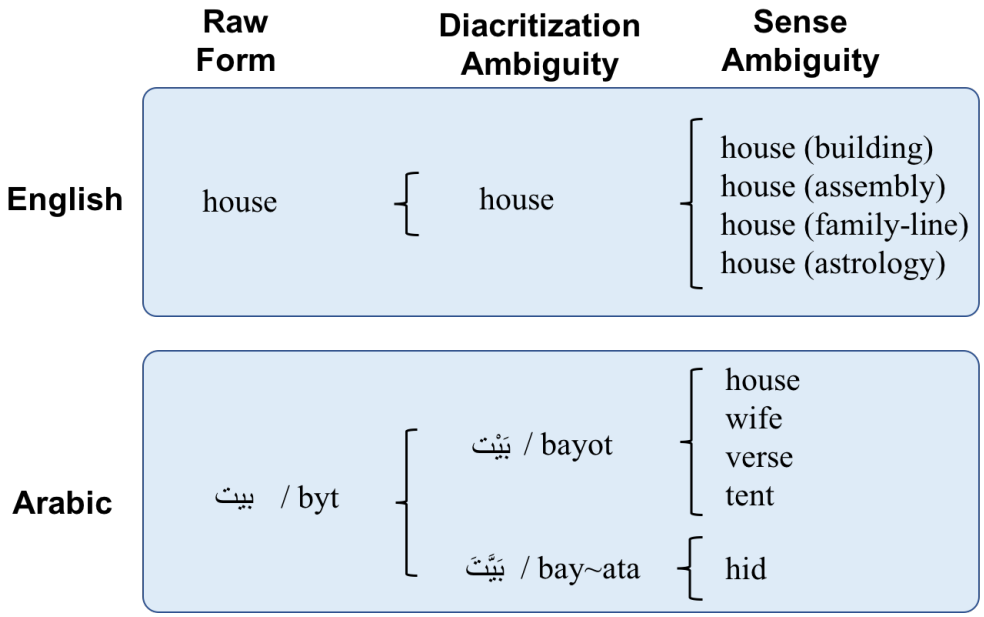



Figure 2: Sources of lexical ambiguity in English and Arabic (from (Hamed and Zesch, 2018)).

situation in English and Arabic. As English uses relatively few diacritics, there is no diacritization ambiguity. For example, the Arabic token بيت /*byt*/ has diacritizations like بَيْت /bayot/ and بَيَّتَ /bay~ata/. As can be seen in the last column in Figure 2, this issue is not to be confused with the sense ambiguity that exists in both English and Arabic on top of the diacritization ambiguity.

Recently, Hamed and Zesch (2017b) have shown that non-diacritized Arabic lexical recognition tests show serious ceiling effects as they are too easy for most learners. It is sufficient for a learner to recognize the root form as they know one of its diacritized forms – probably the most frequent diacritized of a word. Table 1 shows the frequency counts of some diacritized forms of the root /*kr/.[2]

Our hypothesis in this paper is that we can construct a more appropriate Arabic lexical recognition test by using less frequent diacritized forms,

| Surface form | Diacritized form | Gloss | Counts |
|---|---|---|---|
| ذكر | ذَكَر /*~akar/ | Male | 18 |
| | ذِكْر /*ikor/ | Prayer | 10 |
| | ذَكَرَ /*akar/ | He mentioned | 1454 |
| | ذُكِرَ /*ukir/ | It was mentioned | 2001 |
| | ذَكَّرَ /*~akar/ | He reminded | 1 |
| | ذُكِّرَ /*uk~ir/ | He was reminded | 4 |

Table 1: Examples of diacritized forms of the Arabic word ذكر /*kr/.

such as /*ak~ara/ or /*uk~ira/. For that purpose, we first have to find a way to reliably estimate the frequency of diacritized word forms. Then, we conduct a user study, measuring the difficulty of the resulting lexical recognition test under three conditions: (i) No Diacritics: non-diacritized words, (ii) Frequent Diacritics: diacritized using the most frequent diacritized word form, and (iii) Infrequent-Diacritics: diacritized using the least frequent diacritized form of a word.

## 2 Counting Arabic Words

Obtaining reliable frequency counts for Arabic words is a task that entails a lot of NLP challenges regarding availability of corpora, automatic diacritization, segmentation, etc.

[2]The frequency counts are based on the Tashkeela corpus (Zerrouki and Balla, 2017), a corpus of classical Arabic books texts that are provided with diacritics.

| Resource | Proportion |
|---|---|
| Aljazeera online | 30% |
| Arabic Wikipedia | 20% |
| Novels | 15% |
| Alquds newspaper | 10% |
| Altibbi | 10% |
| IslamWeb | 5% |
| Social networks (FB, Twitter) | 5% |
| Other | 5% |

Table 2: Proportion of corpus resource.

### 2.1 Availability of Corpora

We typically need a large amount of diacritized Arabic text to estimate the frequency of diacritized word forms, but there is a lack of such resources. Generally, the currently available diacritized corpora are limited to Classical Arabic (usually religious text), such as the Holy Quran[3], Hadith books, RDI[4] and Tashkeela (Zerrouki and Balla, 2017); or Modern Standard Arabic (usually commercial news wires), such as Penn Arabic Treebanks (ATB) and Agence France Presse (AFP) that can be purchased from the Linguistic Data Consortium (LDC).

**Source Corpus** As the costs of acquiring annotated corpora can prevent researchers from conducting their research, we only want to use freely available corpora. One option is the provided by Zaghouani (2014) and contains newspaper articles crawled from the internet.[5]. However, as we are trying to build an educational application that measures language proficiency, we need text that covers a broader variety of topics. We are thus using the corpus introduced by Freihat et al. (2018), which was assembled from texts and text segments from a varied set of online Arabic language resources such as Wikipedia, news portals, online novels, social media, and medical consultancy web pages. Table 2 shows the distribution of sub corpora in the resource.

### 2.2 Automatic Diacritization

It has been shown that automatic diacritization can be used to obtain reliable frequency counts for Arabic words (Hamed and Zesch, 2018) by automatically diacritizing a large non-diacritized source corpus. According to a recent benchmark (Hamed and Zesch, 2017a) comparing the available tools for diacritization (Farasa (Darwish and Mubarak, 2016), Madamira (Pasha et al., 2014) and two strong baselines), Farasa is outperforming the other approaches under all conditions. Therefore, we use Farasa to diacritize the crawled source corpus. The diacritized corpus is available upon request.

### 2.3 Lemmatization

As we want to use lemmas, not surface forms in our Arabic lexical recognition test, we need to perform lemmatization. This step is necessary as Arabic is a morphology-rich language and its words are highly inflected and derived (Aqel et al., 2015). Darwish and Mubarak (2016) reported that Farasa outperforms or matches state-of-the-art Arabic segmenters/lemmatizers like QCRI Advanced Tools For Arabic (QATARA) (Darwish et al., 2014) and Madamira (Pasha et al., 2014).

We (Hamed and Zesch, 2018) explore the effects of diacritization on Arabic frequency counts. We have shown that Farasa clearly gives better estimates than Madamira. Therefore, we integrate Farasa segmenter/lemmatizer in our NLP pipeline.

### 2.4 NLP Pipeline

To reliably estimate the frequency counts for the diacritized LRT word items, we run the following NLP pipeline, given the source corpus as input: (i) diacritize the source corpus using the Farasa diacritizer, (ii) segment the space-delimited diacritized words using Farasa, (iii) discard the extra clitics, (iv) label the roots with the corresponding diacritics with the help of DKPro Core[6], a collection of software components for natural language processing based on the Apache UIMA framework, and (v) assign the frequency counts for each root based on the attached diacritics.

After carrying out the aforementioned NLP pipeline on this source corpus, we will get frequency counts similar to that in Table 1. The frequency counts contain, among others, the most and least frequent diacritized form of a word that are corresponding to a given non-diacritized root/lemma. Now we are ready to construct the tests and conduct the user study.

[3] http://tanzil.net/download/

[4] http://www.rdi-eg.com/RDI/TrainingData/

[5] Available at: https://sites.google.com/site/mouradabbas9/corpora

[6] https://dkpro.github.io/dkpro-core/

| Word | Nonword | Swapped-letter |
|---|---|---|
| عاقل | عافل | ف to ق |
| رخ | رفخ | خ to ح |
| معكوس | معكوش | ش to س |

Table 3: Nonwords created by letter transposition

## 3 User Study Setup

In order to investigate the role of diacritical marks on improving the construct validity of Arabic lexical recognition tests, we conduct a user study where we compare three tests that differ in the diacritization settings.

- **No Diacritics (S1)**: We use the non-diacritized version of 'test A' as used by Hamed and Zesch (2017b). The nonwords have been generated using a letter substitution/transposition approach in an existing word. Table 3 contains some examples of such nonwords.

- **Frequent-Diacritics (S2)**: We diacritize all roots from S1 with the *most* frequent diacritized form. The nonwords are the same as in S1 and diacritized using a pronounceable (plausible) version of diacritics.

- **Infrequent-Diacritics (S3)**: We diacritize all words from S1 with the *least* frequent diacritized form. Figure 3 shows the resulting test in checklist format.


Figure 3: The diacritized tests items for test A in *infrequent-diacritics* setting (S3), words are checked, nonwords are not.

**Pilot Study** Before conducting the main user study, an Arabic teacher reviewed the three tests. For example, he made sure that no dialectal words are used because they could only be recognized by Arabic speakers of that dialect.

A few students (n = 11) were asked to participate in the user study, so that we check the overall format, design, and test instructions. No modifications have been made to overall test format or design. Minor modifications had to be made to test instructions after the pilot study.

**Main Study** First, we provide participants with a set of instructions including some sample items. Then the participants were asked to provide information about gender, age, mother tongue (L1), and the knowledge of Arabic language (number of years they had taken Arabic courses). Then, participants had to finish the actual lexical recognition test. The test version which participants received (non diacritics, frequent diacritics, infrequent diacritics) was assigned randomly to avoid sequence effects.

**Web Interface** In order to conduct the study, we created a multi-device web interface using PHP and MySql database. Figure 4 shows how it looks like. We make the implementation available to allow for easy replication.[7]

[7] https://github.com/ohamed/ar-lrts

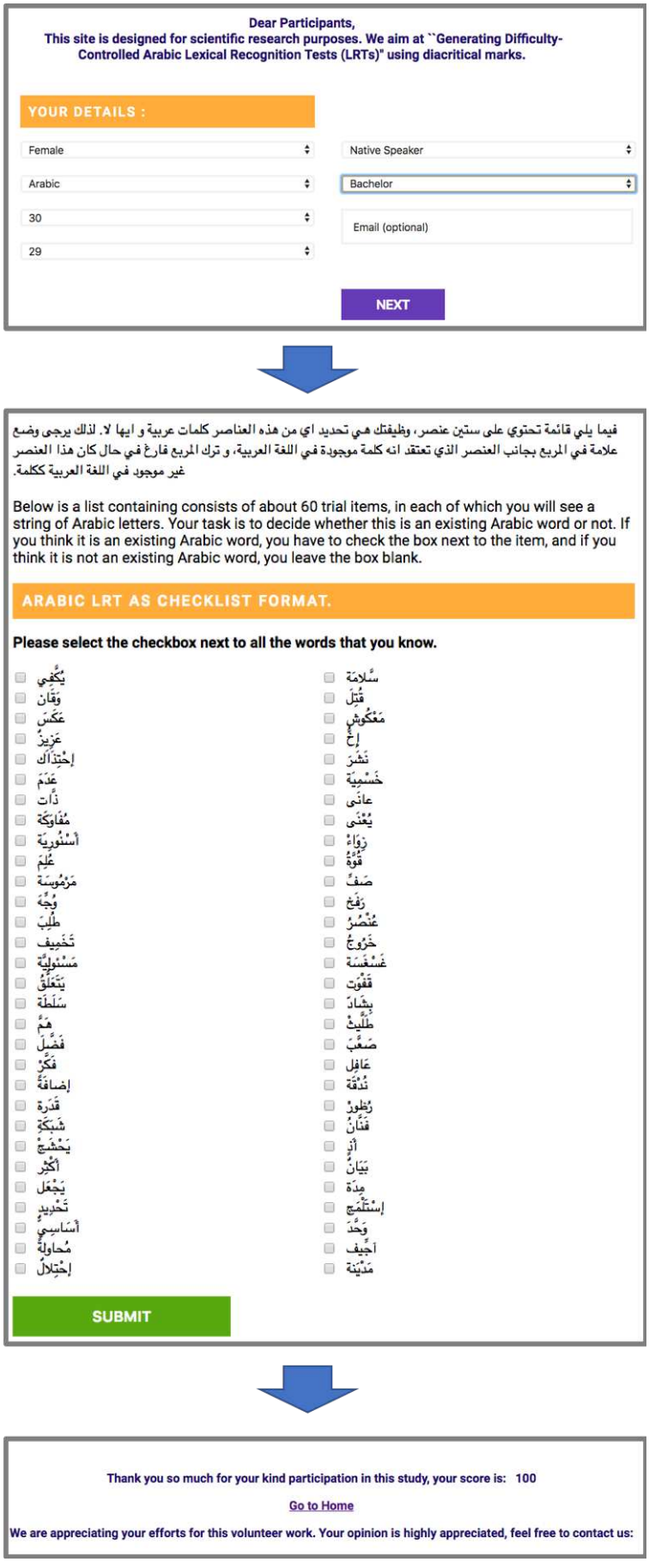



Figure 4: Web system.

| Test Setting | 40 Words | | | 20 Nonwords | | |
|---|---|---|---|---|---|---|
| | P | R | F | P | R | F |
| S1 – No Diacritics | .95 | .95 | .95 | .93 | .89 | .91 |
| S2 – Freq. Diac. | .91 | .92 | .91 | .90 | .82 | .86 |
| S3 – Infreq. Diac. | .92 | .80 | .86 | .71 | .85 | .77 |

Table 4: Results for the three tests settings.

## 4 User Study Results

We advertised our study through different channels, such as mail listings and social media. Overall, 263 people participated in the study, 143 are male, 120 are female. The average age is 28.1 years. Overall, the participants are randomly distributed over the three tests as follows: 96 participants were assigned to S1, 78 participants were assigned to S2, and the remaining 89 participants were assigned to S3.

In Table 4, we show precision, recall, and F-measure for the three test settings for both words and nonwords, averaged over all participants. We see that while the precision for words is comparable over all three tests, our test version S3 with infrequent diacritics has lower recall. This is the intended effect or more people not recognizing the words (remember that the non-diacritized tests are too easy and we want people to fail a bit more often).

### 4.1 Comparing Test Versions

In order to compare the difficulty of the two diacritized tests S2 and S3 with the original non-diacritized test S1, we compute for each respondent a combined test score using the scoring scheme utilized by Hamed and Zesch (2017b). In order to account for the unequal number of words and nonwords in the test, it averages the corresponding recalls.

$$score(R) = \frac{(R_w + R_{nw}) \cdot 100}{2} \qquad (1)$$

This way, a yes bias – by identifying all items as words – (creating high error rates in the nonwords) would be *penalized* in the same way as a no bias – by identifying all items as nonwords – (causing high error rates for words), independently of the different numbers of words versus nonwords.

Then, we compute the average score (over all participants) for each variant. We obtain average scores of 91.8, 86.8, and 82.3 for the three tests respectively. We compute the statistical significance of the differences between the three tests using the *t-test*. All differences between the scores are statistically significant.

We visualize the relationship between the setting and the scores obtained by the participants in each test as shown in Figure 5. The non-diacritized test S1 shows the predicted ceiling effect. The differences to the diacritized version with the most frequent diacritics (S2) are actually larger than we would have predicted (recall that our hypothesis was that even in the non-diacritized version, subjects would fall back to the most frequent diacritized form). However, in line with our predictions the third test version (S3) using infrequent diacritics is much more difficult than both other tests and shows no ceiling effects. It should thus be better suited for accurately measuring the vocabulary size of more advanced learners than the other test versions.

### 4.2 Item Analysis

So far, we have only looked at the test results in general (across all items), but it remains unclear whether all words get more difficult or whether the effect is stronger for some words.

Thus, we visualize the scores for each word in our three experimental settings using a heatmap along with their frequency counts as shown in Table 5. As the score corresponds to how many participants of our study recognized a word, light colors mean easy items and darker colors mean difficult items. We find that some words get much harder when using the least frequent diacritization, while there is almost no effect for other words. In order to check whether this effect can be attributed to the frequency of the underlying forms, we also plot the counts as obtained from the source corpus for the majority of the word items.[8]

Overall, there is no obvious relationship between the scores of the word in the three settings and their frequency counts. For example, هم /hm/ from S1 occurs 4,510 times, هُمْ /humo/ (meaning: *they*) from S2 occurs 2,388 times, and هَمّ /ham∼/ (meaning: *worry*) from S3 occurs 57 times. How-

[8]The frequencies are obtained from the source corpus.

| Arabic | Buckwalter Transliteration | S1 No Diac | S2 Freq. Diac | S3 Infreq. Diac | *freq* S1 | *freq* S2 | *freq* S3 |
|---|---|---|---|---|---|---|---|
| عنصر | EnSr | .99 | .91 | .83 | 50 | 35 | 15 |
| قتل | qtl | .98 | .95 | .95 | 416 | 184 | 77 |
| قوة | qwp | .98 | .92 | .92 | 181 | 115 | 8 |
| صعب | SEb | .98 | .92 | .84 | 132 | 41 | 1 |
| أكثر | Okvr | .98 | .95 | .90 | 1561 | 1120 | 122 |
| أساسي | OsAsy | .98 | .95 | .91 | 753 | 195 | 20 |
| مدينة | mdynp | .98 | .95 | .84 | 98 | 80 | 2 |
| يكفي | ykfy | .97 | .94 | .58 | 139 | 97 | 6 |
| عكس | Eks | .97 | .88 | .90 | 101 | 99 | 2 |
| نشر | n$r | .97 | .90 | .86 | 424 | 181 | 100 |
| عدم | Edm | .97 | .95 | .91 | 931 | 640 | 133 |
| طلب | Tlb | .97 | .94 | .89 | 399 | 192 | 7 |
| خروج | xrwj | .97 | .92 | .68 | 481 | 158 | 21 |
| فضل | fDl | .97 | .92 | .86 | 113 | 84 | 8 |
| فكر | fkr | .97 | .95 | .85 | 332 | 305 | 12 |
| قدرة | qdrp | .97 | .95 | .51 | 34 | 25 | 6 |
| بيان | byAn | .97 | .91 | .91 | 883 | 370 | 3 |
| يجعل | yjEl | .97 | .94 | .90 | 122 | 111 | 11 |
| تحديد | tHdyd | .97 | .94 | .91 | 512 | 310 | 49 |
| سلامة | slAmp | .96 | .96 | .66 | 34 | 26 | 6 |
| عزيز | Ezyz | .96 | .94 | .92 | 472 | 304 | 42 |
| علم | Elm | .96 | .92 | .92 | 348 | 279 | 4 |
| صف | Sf | .96 | .87 | .70 | 131 | 38 | 9 |
| وجه | wjh | .96 | .92 | .80 | 568 | 274 | 12 |
| يتعلق | ytElq | .96 | .90 | .89 | 127 | 110 | 17 |
| شبكة | $bkp | .96 | .91 | .81 | 22 | 19 | 1 |
| محاولة | mHAwlp | .96 | .94 | .92 | 15 | 13 | 2 |
| ذات | *At | .95 | .87 | .65 | 1234 | 205 | 42 |
| إذ | I* | .95 | .91 | .31 | 328 | 302 | 11 |
| مسؤولية | msWwlyp | .94 | .94 | .72 | 734 | 540 | 27 |
| سلطة | slTp | .94 | .91 | .85 | 33 | 27 | 4 |
| هم | hm | .94 | .90 | .93 | 4510 | 2388 | 57 |
| إضافة | IDAfp | .94 | .94 | .91 | 325 | 197 | 5 |
| مدة | mdp | .94 | .95 | .41 | 129 | 92 | 10 |
| أخ | Ox | .93 | .85 | .25 | 38 | 33 | 5 |
| يعني | yEny | .93 | .91 | .86 | 338 | 337 | 1 |
| فنان | fnAn | .93 | .87 | .89 | 876 | 481 | 12 |
| إحتلال | IHtlAl | .93 | .90 | .90 | 316 | 249 | 26 |
| عانى | EAnY | .87 | .87 | .71 | 21 | 14 | 4 |
| وحد | wHd | .65 | .78 | .86 | 335 | 326 | 5 |

Table 5: Heatmap visualizing the average score per word, along with their frequency counts. Items are sorted by S1 score.

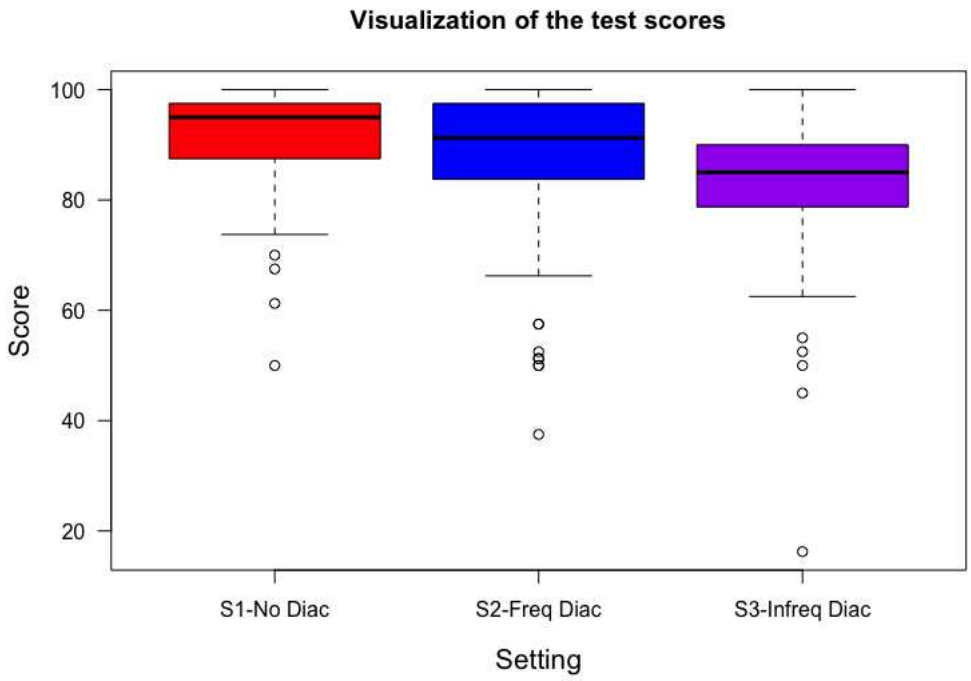



Figure 5: Visualization of the test scores under the three settings.

ever, we don't observe a big drop in the respective scores that are 94%, 93% and 90% for S1, S2, and S3.

## 5 Conclusion & Future Work

In this paper, we have shown that using Arabic lexical recognition tests with less frequent diacritized forms is a way to avoid the ceiling effects of previously proposed non-diacritized tests. We also show how the necessary frequency counts can be obtained by automatically diacritizing source corpora. In future work, we need to further investigate why some infrequent diacritized forms are hard while other (similarly infrequent) diacritized forms are easy. We hypothesize that the corpora used in this study might not reliably reflect the knowledge of learners. Also, even if we tried to minimize the effects of dialects, there might be strong influences from words being frequently used in a dialect or not.

## Acknowledgments

We would like to thank *Andrea Horbach*, the Arabic teacher in the city of Duisburg *Mhamed ben Said*, and my colleagues from the *INDUS network*.

## References

Afnan Aqel, Sahar Alwadei, and Mohammad Dahab. 2015. Building an Arabic Words Generator. *International Journal of Computer Applications*, 112(14).

Harun Baharudin, Zawawi Ismail, Adelina Asmawi, and Normala Baharuddin. 2014. TAV of Arabic language measurement. *Mediterranean Journal of Social Sciences*, 5(20):2402.

Marc Brysbaert. 2013. LEXTALE_FR: A fast, free, and efficient test to measure language proficiency in French. *Psychologica Belgica*, 53(1):23–37.

Kareem Darwish, Ahmed Abdelali, and Hamdy Mubarak. 2014. Using Stem-Templates to Improve Arabic POS and Gender/Number Tagging. In *LREC*, pages 2926–2931.

Kareem Darwish and Hamdy Mubarak. 2016. Farasa: A New Fast and Accurate Arabic Word Segmenter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Ali Farghaly and Khaled Shaalan. 2009. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4):14.

Abed Alhakim Ali Kayed Freihat, Gabor Bella, Mubarak Hamdy, Fausto Giunchiglia, et al. 2018. A single-model approach for arabic segmentation,

pos-tagging and named entity recognition. In *International Conference on Natural Language and Speech Processing ICNLSP 2018*, Algiers, Algeria. ICNLSP.

Nizar Habash. 2010. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Osama Hamed and Torsten Zesch. 2015. Generating Nonwords for Vocabulary Proficiency Testing. In *Proceeding of the 7th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 473–477, Pozna, Poland.

Osama Hamed and Torsten Zesch. 2017a. A Survey and Comparative Study of Arabic Diacritization Tools. *JLCL: Special Issue - NLP for Perso-Arabic Alphabets.*, 32(1):27–47.

Osama Hamed and Torsten Zesch. 2017b. The Role of Diacritics in Designing Lexical Recognition Tests for Arabic. In *3rd International Conference on Arabic Computational Linguistics (ACLing 2017)*, Dubai, UAE. Elsevier.

Osama Hamed and Torsten Zesch. 2018. Exploring the Effects of Diacritization on Arabic Frequency Counts. In *Proceeding of the 2nd International Conference on Natural Language and Speech Processing (ICNLSP 2018)*, Algiers, Algeria.

Ineke Huibregtse, Wilfried Admiraal, and Paul Meara. 2002. Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language testing*, 19(3):227–245.

Cristina Izura, Fernando Cuetos, and Marc Brysbaert. 2014. Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica*, 35(1):49–66.

Kristin Lemhöfer and Mirjam Broersma. 2012. Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, 44(2):325–343.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *LREC*, pages 1094–1101.

Robert Ricks. 2015. The Development of Frequency-Based Assessments of Vocabulary Breadth and Depth for L2 Arabic.

Raymond Stubbe. 2012. Do pseudoword false alarm rates and overestimation rates in yes/no vocabulary tests change with japanese university students english ability levels? *Language Testing*, 29(4):471–488.

Wajdi Zaghouani. 2014. Critical survey of the freely available Arabic corpora. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC'2014), OSACT Workshop. Rejkavik, Iceland*.

Taha Zerrouki and Amar Balla. 2017. Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems. *Data in Brief*, 11:147–151.