Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2017) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5–6 November 2017, Dubai, United Arab Emirates

The Role of Diacritics in Designing Lexical Recognition Tests for Arabic

Osama Hamed*, Torsten Zesch

Language Technology Lab, Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen, 47057 Duisburg, Germany

{*osama.hamed, torsten.zesch*}*@uni-due.de*

Abstract

Lexical recognition tests are widely used to assess vocabulary knowledge. We investigate the role that diacritics play in designing an Arabic lexical recognition test. We compare a non-diacritized and a diacritized test in a user study and find that they are largely comparable in their ability to assess vocabulary proficiency. However, we argue that diacritized tests are better suited to control the test difficulty by allowing better nonwords and a more targeted selection of word forms.

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.

Keywords: Vocabulary Size; Lexical Recognition Tests; Diacritization

1. Introduction

Measuring language proficiency is an important task for educators. Vocabulary size is one part of the overall proficiency that can be used to approximate learning progress or to select a suitable language course [24]. Increasing your vocabulary is an essential component of language learning and is also one of the main conditions in mastering a language [8].

Lexical recognition tests (LRTs) are one of the the best-known and most widely used vocabulary assessment formats [31]. The main advantage is their simplicity [25], as participants are just being shown a list of words and asked to say 'Yes' when they know the word or 'No' otherwise. In order to make the task difficult and to avoid cheating, besides real words like *obey*, also nonwords like *nonagrate* are shown. Figure 1 shows an example for such a test in checklist format. As all words are checked and all nonwords are not checked, this shows a result for a learner with a good vocabulary knowledge.

* Osama Hamed. Tel.: +49 203 379 1433 ; fax: +49 203 379 3557.
E-mail address: osama.hamed@stud.uni-due.de

Peer-review under responsibility of the scientific committee of the 3rd International Conference on Arabic Computational Linguistics.

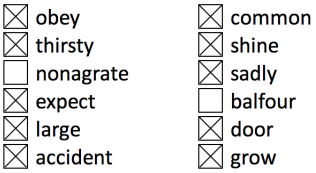


Figure 1: Lexical recognition test in checklist format.

While there exist established lexical recognition tests for English, e.g. LexTALE [22], for many under-resourced languages, like Arabic, a lot of challenges still remain. In this paper, we address some of these challenges by taking a closer look on the design process of Arabic tests and especially the role of diacritical marks that are a defining feature of Arabic. For that purpose, we conduct a user study that compares an existing non-diacritized test for Arabic [29] with an adapted version including diacritics. We also discuss the NLP-related challenges when aiming to automatically build LRTs with diacritics.

Before we go into the details of the study, we provide some background on lexical recognition tests in general and the special challenges faced when creating such tests for Arabic.

2. Lexical Recognition Tests

Lexical recognition tests are used for measuring the vocabulary size of learners [13]. They are based on the assumption that recognizing a word is sufficient for 'knowing' the word, i.e. they only measure the size or breadth of vocabulary knowledge, but not the depth or quality [3, 32]. However, for many purposes like placement tests or quickly assessing the progress of vocabulary acquisition, lexical recognition tests have been successfully applied.

In order to cover a wide difficulty range, the words to be used in such tests are usually selected based on corpus frequency. The tests additionally use carefully selected nonwords that act as distractors. This is necessary, as otherwise learners could easily game the test by simply pretending to 'know' all the words. In a test with a mix of words and nonwords, such a strategy leads to a rather low score.

Lexical recognition tests already achieve a quite good approximation of a learner's vocabulary with a relatively small number of test items [20]. Thus, lexical recognition tests can be quickly finished and usually fit on a single sheet of paper. This is the so called *checklist* format as shown in Figure 1. When used in a computerized form, individual items are usually shown in isolation in order to minimize context effects.

2.1. Existing Test

We now look into previous work on lexical recognition tests, which has mainly focused on English and a few other European languages [29], while very few studies investigated Arabic.

English LRTs. An early example of using nonwords for testing is the *Eurocentres Vocabulary Size Test* [26]. It uses 150 items – two thirds real words that were selected by frequency, and one third manually-crafted nonwords. Lemhöfer and Broersma [22] constructed a smaller version of this test called *LexTALE* that can be finished faster. It only uses 40 words (selected by relative frequency in the CELEX corpus [7]) and 20 manually-crafted nonwords. LexTALE scores are validated by correlating them with other proficiency scores based on a word translation task and the commercial 'Quick Placement Test'. LexTALE has been adapted to other languages beyond English, e.g. Dutch and German [22], French [11], and Spanish [21].

Arabic LRTs. We are only aware of a very limited set of studies on Arabic lexical recognition tests which all use non-diacritized Arabic. Baharudin et al. [9] developed the *Test of Arabic Vocabulary* that uses 40 words selected by corpus frequency, but no nonwords. Thus, the test is vulnerable to test-wiseness or overconfidence (just answering 'yes' for each item).

w/o diacritics شرب الولد حليبا

w/ diacritics شَرِبَ الوَلَدُ حَلِيباً

Figure 2: Example sentence with and without diacritics (eng: *The boy drank milk*).

	/E□l□m□/	Gloss	Count Tashkeela
عِلْم	/Eilom/	Science	4
عَلَم	/Ealam/	Flag	1
عَلِمَ /Elm/ علم	/Ealima/	He knew	792
عُلِمَ	/Eulima/	It was known	433
عَلَّمَ	/Eal~ama/	He taught	18

Figure 3: Examples of diacritized forms of the Arabic word علم /Elm/. Frequency counts are based on the Tashkeela corpus.

Ricks [29] developed a checklist-format test with 40 words and 20 nonwords (following the format introduced with LexTALE). Words were randomly selected from the Buckwalter/Parkinson frequency dictionary [12], but excluding dialectal words. Nonwords were created using letters substitution approach as inspired by [33].

2.2. Generating LRTs

The automatic generation of LRTs involves two steps: (i) selecting words from a corpus and (ii) generating nonwords. In the past, nonwords have been manually created [22]. However, for the repetitive testing as used in formative assessment [34], nonwords test stimuli need to be generated automatically. Hamed and Zesch [19] proposed an approach to generate nonwords automatically using character n-gram language models. They applied their approach to English, and considered word selection using frequency per million word.

3. Role of Diacritics in Arabic LRTs

We now turn to the role of diacritics in Arabic lexical recognition tests. We argue that they play a crucial role in selecting words as well as designing suitable nonwords.

Usually, learners of Arabic build their vocabulary knowledge from diacritized material. Especially in the early stages, learners might find it difficult or unnatural to recognize words without diacritics. For example, all the textbooks in the series "I Love the Arabic Language" are diacritized.[1] Thus, a non-diacritized lexical recognition test might systematically underestimate the performance of low proficiency learners.

Selecting Words. Arabic text is mostly written non-diacritized, i.e. without any diacritical marks, except for religious texts, educational texts, and some poetry [17]. Figure 2 shows the non-diacritized and the diacritized versions of the sentence "*The boy drank milk*".

The analogy with English is imperfect, but in a sense the situation would be similar to presenting someone the string *str* and expecting them to be able to determine whether the intended English word is *star*, *stir*, *suitor*, *sitar*, or *store* depending on the context [30]. So, in a lexical recognition test, when we ask a learner if she knows the 'word' *str*, we are actually asking whether she knows any of the words from the list above which is quite an imprecise question.

The lack of diacritics usually leads to considerable lexical and morphological ambiguity [35]. Following an example from Maamouri et al. [23], we show in Figure 3 a non-exhaustive list of diacritizations of the Arabic word علم

[1] http://www.noorart.com/school_section/i_love_the_arabic_language_arabic_curriculum

/Elm/. We hypothesize that the difficulty to recognize a non-diacritized word is actually determined by the relative probability of its most-frequent diacritized form. We also argue that this effect goes way beyond the related issue of word senses for English lexical recognition tests, where showing a word like *tree* also tests whether one knows the most frequent sense (*a tall perennial woody plant*) and not one of the more specialized ones (*data structure in computer science*). As we can see in Figure 3, the ambiguity introduced by non-diacritized text includes phonological, morphological, and syntactic cases [35].

In order to determine the most likely diacritization, we can check the frequency based on diacritized corpora. We use a subset of 11 books of the Tashkeela corpus [36]. Because of the religious nature of the texts in this corpus, the counts for *Science* and *Flag* are very low, while *He knew* and *It was known* are two orders of magnitude higher. Thus, learners of Arabic that mainly read religious texts will be able to recognize the diacritized form of /Elm/ meaning *He knew*, but not *Science*, while no such distinction can be made in the non-diacritized version of the test.

Designing Nonwords. So far we have only discussed existing word forms, but diacritics might also play a crucial role in designing better nonwords. Arabic diacritization is an orthographic way to describe Arabic word pronunciation [35]. We hypothesize that the non-diacritized nonwords are probably easier than the diacritized ones. The diacritized nonwords can distract better with closely related, especially if they are labeled with pronounceable diacritics.

4. Constructing a Diacritized Arabic Test

We now discuss the linguistic and technical challenges that occur in the two steps (word selection and nonword generation) of automatically constructing a lexical recognition test.

4.1. Selecting Arabic Words

Selecting words for lexical recognition tests is mainly based on frequency counts. However, obtaining reliable frequency counts is more challenging in Arabic than in English due to complex morphology and the issue of diacritization.

Morphology. Arabic is a morphology rich language and its words are highly inflected and derived [5]. For example, the word (“wktAbnA”, ‘وكتابنا’, and our book) consists of three clitics “w+ktAb+nA”: (i) the conjunction article “w” as prefix, (ii) the stem “ktAb”, and (iii) the possessive pronoun “nA” as suffix.

So in order to get a reliable frequency count for the lemma *ktAb* (engl. *book*), we have to use segmenters and lemmatizers to discard such extra clitics [17]. Fortunately, there are tools such as Farasa [14] that are specifically designed for that task.

Another example of morphology standing in the way of frequency counting is the pervasiveness of the definite article “Al” (ال) that is directly attached to a word. For example, in the arTenTen corpus [10], which comprises 5.8 billion words, the frequency of the word (“Tfl”, ‘طفل’, child) is 4,557, whereas the frequency of the same word along with the definite article (“AlTfl”, ‘الطفل’, the child) is 15,325.

Automatic Diacritization. We suggest using the diacritized lemmas for a better frequency count. However, as there is only a limited number of rather small corpora with manually annotated diacritics [2] one has to fall back to automatic diacritization in order to obtain reliable frequency counts of diacritized word forms.

Although there is a large body of research on the topic [6, 27], only very few tools are freely available and it is still unclear what performance level can be expected in a practical setting. It is especially unclear whether existing tools will simply project the distribution of diacritics found in the training corpus to new data or if they generalize well enough to be useful for the purposes of designing lexical recognition tests. We are not aware of any research that actively targets this question.

4.2. Designing Arabic Nonwords

In a lexical recognition test, a good nonword acts as a distractor, i.e. it is similar enough to real words that it forces test-takers to be careful about their answers. However, nonwords should of course not be a valid word from the vocabulary of a language.

MSA	Dialectal	Nonword
حريق	حريء	هريء

Table 1: Nonwords and diglossia.

Anderson and Freebody [4] discuss two methods for creating nonwords in English: (i) pseudo-derivatives, which entails adding a prefix or suffix to a real word, so 'loyal' becomes 'loyalment'; and (ii) letter substitution, where one or two vowels and/or consonants are substituted in a real word, so 'boy' becomes 'poy'. However, substituting letters is complicated due to the peculiarities of the Arabic alphabet.

Dealing with Similar Shapes. There are two types of symbols in the Arabic script for writing words: letters and diacritics [18]. Arabic letters typically consist of two parts: letter form and letter mark [17]. The Arabic alphabet has a total of 19 letter forms. The letter marks fall into three sub-types: the dots, the short Kaf and the Hamza (ء).[2] The Arabic alphabet features a significant number of letters that differ only in the position (e.g. خ ح ج) or number (e.g. ب ث ت) of dots placed around the letter form. Thus, for instance, the nonword "jArx" (جارخ) and the real word ("xArj", 'خارج', outside) are differentiated only by the placement of two points. This is more subtle than, for example, English *fish* versus *shif*.

Dealing with Similar Sounds. An orthography is a specification of how the sounds of a language are mapped to/from a particular script [17]. Some phonemes of Arabic language have emphatic counterparts. The learners in a type of phonemic L1 transfer, have a tendency to conflate these L2 phoneme pairs. Special attention should be paid to ensure that nonwords could not be coined, by means of substituting one or more of these "confusable" pairs of letters with similar sounds like ت /t/ and ط /T/ or د /d/ and ض /D/.[3]

One nonword, for example, that is hard to be rejected is "ItfAl"[4] (إتفال), as it can be easily confused with the real word ("⟩TfAl" or "OTfAl", 'أطفال', children). Another good example for hard rejection is "DfDE" ضفضع, which should be easily confused with the real word ("dfDE", 'ضفدع', frog).

Dealing with Diglossia. The Arabic language has at least three forms: Classical Arabic, Modern Standard Arabic (MSA), and Dialectal Arabic [15]. This leads to situations where a speaker of Arabic might use two varieties of the language. This kind of situation is what is linguistically known as diglossia [16].

Consider, for example, the MSA word (*Hryq*, 'حريق', fire) and the dialectal word (*Hry'* حريء) which is the Syro-Lebanese counterpart of the MSA word. As nonwords are often generated by means of swapping one letter, *hry'* (هريء) could be generated by swapping the ح /H/ with ه /h/ in Syro-Lebanese instance. The three instances are shown together in Table 1. However, this is problematic, as the dialectal word (*Hry'* حريء) is well known in the Levantine area. Thus such a nonword would be much easier for Syro-Lebanese speakers and much more difficult for others, as it is much closer to an existing word than when looking at MSA only.

The same argument can be extended to real words. For example, the MSA word ("jbhp", 'جبهة', forehead) has many dialectal counterparts (a subset is shown in Table 2). The MSA word is rather similar to the word in the Gulf dialect, while it would be more difficult to recognize for speakers of Egyptian and Maghreb dialects whose dialectal words are quite different.

5. User Study

In order to investigate the role of diacritical marks on Arabic lexical recognition tests, we conduct a user study where we compare a non-diacritized and a diacritized test. In order to avoid memorization effects, one student cannot

[2] The Letter marks, specifically the dots, should not be confused with Hebrew Niqqud 'dots', which are optional diacritics comparable to Arabic diacritics. Arabic dots and other letter marks are all obligatory [17].

[3] In almost cases, the capital letters are used to refer to a stressed letter.

[4] Transliterated using Safe Buckwalter.

MSA	Gulf	Egyptian	Maghreb
جبهة	ييبهة	قورة	فرنت

Table 2: The dialectal varieties for MSA word "jbhp" (جبهة).

(a) test A (b) test B

Figure 4: The two diacritized tests used in our study. Words are checked, nonwords are not.

answer the non-diacritized (ND) and diacritized (D) version of the same test. Thus, in our user study, we use two tests of ND/D pairs. In order to avoid sequence effects, one group begins with the diacritized version, while the other group begins with the non-diacritized one. We visualize this setup in the following figure:

Group 1 (G1)	Group 2 (G2)
test A (D)	test A (ND)
test B (ND)	test B (D)

As a starting point, we utilize two non-diacritized Arabic tests prepared by Ricks [29] that both contain the traditional number of 40 words and 20 nonwords. In order to avoid guessing, the items were randomized.[5] We created a diacritized version of both tests by using the most probable form (see discussion in Section 3 and especially Figure 3). For nonwords, we use a plausible version of diacritics. We also normalized all the initial Alif letters by adding the Hamza (glottal stop) to both versions of the test. Figure 4 shows the diacritized test versions.

We provided the participants in our study with a set of instructions including some sample items. The study itself was implemented as a paper-based survey under direct supervision of a teacher. We recruited 40 students (22 female) from 5 German schools (4^{th} to 10^{th} grade), who are studying the Arabic European syllabus "I Love the Arabic Language" that conforms to the Common European Framework of Reference (CEFR).[6] All students are native German speakers, but with Arabic as a family language.

[5] In the original version, all the words were presented first and the nonwords after that. This is clearly not optimal, as participants can quite easily detect and exploit this setup.

[6] http://www.englishprofile.org/index.php/the-cef

Students	Words						Nonwords					
	ND			D			ND			D		
Group	P	R	F	P	R	F	P	R	F	P	R	F
G1	.93	.70	.78	.96	.75	.82	. 66	.91	.75	.71	.94	.79
G2	.95	.81	.86	.96	.85	.89	.77	.93	.82	.81	.93	.86

Table 3: Results for non-diacritized and diacritized tests comparing groups.

In order to provide an external gold standard for the proficiency level of each student, we asked the Arabic teacher (*before* conducting the study) to evaluate each student on a three-point proficiency scale: (1) beginner, (2) intermediate, and (3) advanced. This gold standard provides a basis for judging the construct validity of our Arabic lexical recognition tests [28].

6. Results and Discussion

Our test design that tries to avoid memory effects entails that the non-diacritized (ND) and diacritized (D) variant of a test are solved by different groups. In order to make meaningful comparisons between the ND and D variants, we first have to make sure that the groups are comparable.

6.1. Group comparison

In Table 3, we show precision, recall, and F-measure for both groups. We see that while the precision is comparable for both groups, group 2 has higher recall in general which can be explained by a slightly higher number of high proficiency students in this group. In general, we see that for words, the precision is quite high, while for nonwords the recall is quite high. This is related to the strategy applied by most students that they only check the words that they actually know. Leading to high precision for words, and high recall for nonwords which is the fallback.

Scoring. There are several possible methods to score LRTs. We only want one combined score for word and nonword performance - in order to avoid test-wiseness effects, e.g. students answering that they know all the words. For each participant, we compute the test score using the scoring scheme introduced for LexTALE, as it turned out to yield the best results [22].

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

The score consists of the ratio of correct responses for words and nonwords – i.e. the recall for each class. This way, a yes bias (creating high error rates in the nonwords) would be *penalized* in the same way as a no bias (causing high error rates for words), independently of the different numbers of words versus nonwords.

Figure 5 shows a scatterplot of the two groups regarding the assigned test scores. It also confirms our finding that both groups are comparable.

6.2. Proficiency level

Figure 6 visualizes the relationship between the evaluation of the teacher and the scores assigned by our two test versions. Both versions assign on average higher scores to more proficient students, i.e. they measure the language proficiency to some extent. The non-diacritized (ND) version of the tests has higher variance for the low proficiency students, while the diacritized (D) version has higher variance for the high proficiency students. However, due to our relatively small sample size, we cannot draw definite conclusions from that observation.

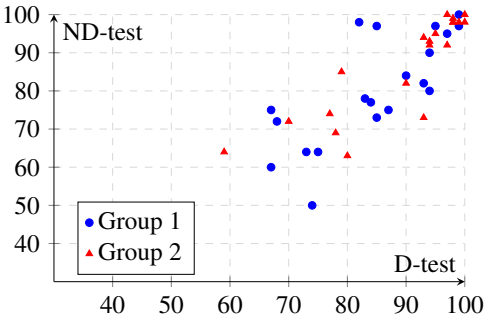
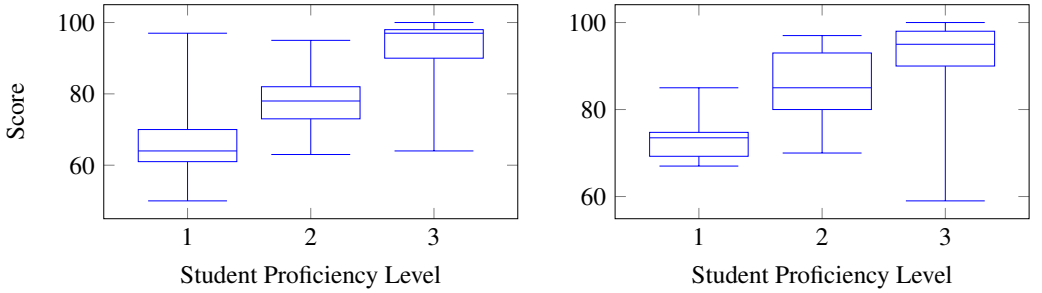


Figure 5: Participants' scores on the ND-test & D-test.



(a) Non-diacritized test (ND)

(b) Diacritized test (D)

Figure 6: Visualization of teacher evaluation in the ND-test and D-test.

Proficiency	Words						Nonwords					
	ND			D			ND			D		
Level	P	R	F	P	R	F	P	R	F	P	R	F
1	.88	.50	.61	.95	.54	.67	.50	.86	.55	.52	.93	.66
2	.93	.68	.77	.96	.78	.85	.62	.90	.68	.71	.93	.80
3	.97	.87	.91	.97	.89	.91	.83	.95	.87	.86	.94	.88

Table 4: Results grouped by proficiency level.

In order to analyze the differences between the three levels, we additionally show a breakdown of precision, recall, and F-measure grouped by proficiency level in Table 4. We observe the usual trend of high precision or words, and high recall for non-words related to the test strategy of only checking known words. We also see that the results for the D and ND variants of our test are relatively similar, which means that using precision instead of recall in Equation 1 would not make much of a difference.

6.3. Qualitative Analysis

Table 5 shows the most difficult pair of word and nonword for each group and test condition. The two words with the highest difficulty index are **aAto* and *IgtyAl*. Both are selected by 5 (i.e. missed by 15) students out of 20. This is most likely because **aAto* normally appears as a noun-phrase, whereas *IgtyAl* is most common in countries with conflicts like Palestinian territories. The two nonwords with the highest difficulty index are *xsmyp* and *mukaAdaAp*. Both nonwords are very similar to Arabic words: (i) *xsmyp* is similar to ("Hsmyp", '**حسمية**', finality) and (ii) *mukaAdaAp* is similar to ("muqaADaAp", '**مُقاضاة**', prosecution). An interesting observation is that non-words have much lower error rate than words. This is mainly due to the above mentioned strategy of only checking words one really knows.

	Words				Nonwords		
Group, Test	**Arabic**	**Transliteration**	**Meaning**	**%Wrong**	**Arabic**	**Transliteration**	**%Wrong**
G2, test A (ND)	وحد	wHd	unify	.55	خسمية	xsmyp	.40
G1, test B (ND)	إغتيال	IgtyAl	assassination	.75	زماء	zmA’	.35
G1, test A (D)	ذَاتْ	*aAto	self	.75	بِشَادّ	bi$aAd	.25
G2, test B (D)	إِغتِيَالْ	IigityaAlo	assassination	.50	مُكَادَاة	mukaAdaAp	.40

Table 5: List of most difficult items for each test variant.

6.4. Limitations

As it can be clearly seen from the results of our study, the scores for the advanced learners are relatively high, i.e. we are already seeing ceiling effects here. As a consequence, in the current form the tests cannot be used for truly advanced students.

Nonwords. As our empirical results show, nonwords generated from non-existing roots are too easy to spot. The same is true for nonwords composed of rare phonemes or rare combinations of phonemes that feature too many of the least common letters, which can cause the stimuli items composed from them to appear unlikely to test respondents including those with beginner or intermediate levels. However, we believe that using diacritics opens the possibility to utilize nonwords that use existing roots, but with a non-existing configuration of diacritics. Exploring this option remains as future work.

Words. Generally, the quality of the words is acceptable, but can be further improved as we are seeing some ceiling effects even for medium proficiency students, i.e. some of the words are much too easy. Consequently, we need a better way of controlling the frequency. The Buckwalter/Parkinson frequency dictionary provides a list of the 5,000 most frequently used words in MSA as well as several of the most widely spoken Arabic dialects. A better valid option might be the revised Arabic WordNet, which comes with irregular plurals [1]. Even better would be corpora reflecting the type of reading material students are likely to have seen at a certain proficiency level, but to the best of our knowledge no such corpus is currently available.

7. Conclusion and Future Work

We have tackled the task of designing lexical recognition tests in Arabic by first discussing the specific challenges that are imposed by the language. It seem clear that using diacritics has the potential to (i) improve the quality of nonwords and (ii) to better control the difficulty of the tests.

We compared the diacritized and non-diacritized lexical recognition tests in a user study and find that they are largely comparable. This is in line with our hypothesis that students will recognize the most probable diacritized word which we used in our test.

In future work, we want to use less likely diacritized forms and explore how well we can control the difficulty of the tests. We envision to create tests that are better able to discriminate medium and high proficiency learners as we already see ceiling effects in the non-diacritized test versions, mainly due to very easy nonwords. We also want to explore ways to automatically create Arabic lexical recognition tests, a task that entails a lot of NLP challenges regarding automatic diacritization, morphological analysis, and language modeling.

References

[1] Abouenour, L., Bouzoubaa, K., Rosso, P., 2013. On the evaluation and improvement of Arabic WordNet coverage and usability. Language resources and evaluation 47, 891–917.

[2] Al-Sulaiti, L., Atwell, E.S., 2006. The design of a corpus of contemporary Arabic. International Journal of Corpus Linguistics 11, 135–171.
[3] Anderson, R.C., Freebody, P., 1981. Vocabulary Knowledge. DOCUMENT RESUME CS 006 138 Guthrie, John T., Ed. Comprehension and Teaching; Research Reviews. International Reading Association, Newark, Del. , 77.
[4] Anderson, R.C., Freebody, P., 1983. Reading comprehension and the assessment and acquisition of word knowledge. Advances in reading/language research .
[5] Aqel, A., Alwadei, S., Dahab, M., 2015. Building an Arabic Words Generator. International Journal of Computer Applications 112.
[6] Azmi, A., Almajed, R., 2015. A survey of automatic Arabic diacritization techniques. Natural Language Engineering 21, 477–495.
[7] Baayen, R.H., Piepenbrock, R., Gulikers, L., 1995. The CELEX lexical database (release 2). Linguistic Data Consortium, Philadelphia .
[8] Baharudin, H., Ismail, Z., 2014. Vocabulary Learning Strategies and Arabic Vocabulary Size among Pre-University Students in Malaysia. International Education Studies 7, 219–226.
[9] Baharudin, H., Ismail, Z., Asmawi, A., Baharuddin, N., 2014. TAV of Arabic language measurement. Mediterranean Journal of Social Sciences 5, 2402.
[10] Belinkov, Y., Habash, N., Kilgarriff, A., Ordan, N., Roth, R., Suchomel, V., 2013. arTen-Ten: a new, vast corpus for Arabic. Proceedings of WACL .
[11] Brysbaert, M., 2013. Lextale_FR a fast, free, and efficient test to measure language proficiency in French. PSYCHOLOGICA BELGICA 53, 23–37.
[12] Buckwalter, T., Parkinson, D., 2011. A frequency dictionary of Arabic: Core vocabulary for learners. Routledge.
[13] Cameron, L., 2002. Measuring vocabulary size in English as an additional language. Language Teaching Research 6, 145–173.
[14] Darwish, K., Mubarak, H., 2016. Farasa: A New Fast and Accurate Arabic Word Segmenter, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Paris, France.
[15] Farghaly, A., Shaalan, K., 2009. Arabic natural language processing: Challenges and solutions. ACM Transactions on Asian Language Information Processing (TALIP) 8, 14.
[16] Ferguson, C.A., 1959. Diglossia. word 15, 325–340.
[17] Habash, N., 2010. Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies 3, 1–187.
[18] Habash, N., Rambow, O., 2007. Arabic diacritization through full morphological tagging, in: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers, Association for Computational Linguistics. pp. 53–56.
[19] Hamed, O., Zesch, T., 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, Poznań, Poland. pp. 473–477. URL: http://www.ltl.uni-due.de/wp-content/uploads/nonwords-ltc20151.pdf.
[20] Huibregtse, I., Admiraal, W., Meara, P., 2002. Scores on a yes-no vocabulary test: Correction for guessing and response style. Language testing 19, 227–245.
[21] Izura, C., Cuetos, F., Brysbaert, M., 2014. Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. Psicológica 35, 49–66.
[22] Lemhöfer, K., Broersma, M., 2012. Introducing LexTALE: A quick and valid lexical test for advanced learners of English. Behavior Research Methods 44, 325–343.
[23] Maamouri, M., Bies, A., Kulick, S., 2006. Diacritization: A challenge to Arabic treebank annotation and parsing, in: Proceedings of the Conference of the Machine Translation SIG of the British Computer Society, Citeseer.
[24] Maskor, Z.M., Baharudin, H., Lubis, M.A., Yusuf, N.K., 2016. Teaching and Learning Arabic Vocabulary: From a Teacher's Experiences. Creative Education 7, 482.
[25] Meara, P., Jones, G., 1987. Tests of vocabulary size in English as a foreign language. Polyglot 8, 1–40.
[26] Meara, P., Jones, G., 1990. Eurocentres vocabulary size test 10KA. Zurich: Eurocentres .
[27] Metwally, A.S., Rashwan, M.A., Atiya, A.F., 2016. A multi-layered approach for Arabic text diacritization, in: Cloud Computing and Big Data Analysis (ICCCBDA), 2016 IEEE International Conference on, IEEE. pp. 389–393.
[28] Milton, J., 2007. Lexical profiles, learning styles and the construct validity of lexical size tests. Modelling and assessing vocabulary knowledge , 47–58.
[29] Ricks, R., 2015. The Development of Frequency-Based Assessments of Vocabulary Breadth and Depth for L2 Arabic .
[30] Saigh, K., Schmitt, N., 2012. Difficulties with vocabulary word form: The case of Arabic ESL learners. System 40, 24–36.
[31] Schmitt, N., 2000. Vocabulary in language teaching. Ernst Klett Sprachen.
[32] Schmitt, N., 2014. Size and depth of vocabulary knowledge: What the research shows. Language Learning 64, 913–951.
[33] Stubbe, R., 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, 471–488.
[34] Wang, T.H., 2007. What strategies are effective for formative assessment in an e-learning environment? Journal of Computer Assisted Learning 23, 171–186.
[35] Zaghouani, W., Bouamor, H., Hawwari, A., Diab, M., Obeid, O., Ghoneim, M., Alqahtani, S., Oflazer, K., 2016. Guidelines and framework for a large scale Arabic diacritized corpus, in: The Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 3637–3643.
[36] Zerrouki, T., Balla, A., 2017. Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems. Data in Brief 11, 147–151.