**ALQUDS UNIVERSITY**



**FACULTY OF SCIENCE & TECHNOLOGY**

**DEAPRTMENT OF COMPUTER SCIENCE**

**MASTER’S IN COMPUTER SCIENCE**

**THESIS PROPOSAL**

**Towards the Automatic Generation of Diacritized Arabic Lexical Recognition Tests**

**Mohammad Nassar**

[**nassar.mohammad@students.alquds.edu**](mailto:nassar.mohammad@students.alquds.edu)

**Mar 06, 2018**

**EXECUTIVE SUMMARY**

In this study, we proposed construction of Lexical Recognition Test for Arabic language LRT, LRT exam will include diacrtizaed words and non-words, student will determine the real word vocabulary from a checklist or true/false list.

LRT based on the hypothesize that recognizing a word means knowing that word, dialects words are out of study scope, only root words of modern standard language will be included.

1. **1. THE TOPIC | MOTIVATION | PROBLEM**

Nowadays many foreign people pay attention to learn Arabic language, Muslims who don’t use Arabic as the mother language need it to read the holy Quran and know more about Islam, So many researcher proposed an approach to measure Arabic proficiency of the foreign learners with a simple, and fast placement test that take about a few minutes, of course the existing research for English language and some other languages like German is rich, but unfortunately the research about the Arabic still scarce.

*Key words: NLP, Diacritized, corpus, Lexical recognition test (LRT), Vocabulary Size, non-words, n- gram, tokenization, crawling, stimuli, corpora, Lemmatization.*

**Our research motivation** inherited from multi-reasons as Arabic language considered one of the top seven languages that people communicate through around the world, in Arab world only there is more than 400 million person who use Arabic as mother tongue language; besides this we can notice high interest for understanding Arabic reviews over the web and social media to understand Arabs express towards situations, finally Muslims around the world need to know better about their religion so they have to obtain some level of Arabic.

**Current Problem** could be noticed since learner can’t find an effective measurement to indicate their level of recognition and knowledge in Arabic language.

Based on current situation, this research contribution aims at developing Arabic lexical recognition tests automatically by means of using a standard language model proved through implementing a web based application that generates placement test to measure learner level. This will help the learners in determining their proficiency level in Arabic.

Lexical recognition test (LRT) is a vocabulary size test. LRTs are frequently used for measuring language proficiency. In such test, students are typically shown either valid words like “denial” or non-words like “platery”, and they need to decide if they are valid or not. The main advantage of a LRT is its simplicity. It only takes five minutes, only ``Yes/No'' or checklist questions are asked - as shown in Figure 1, and scoring can easily be automated.

|  |
| --- |
| Macintosh HD:Users:ltl2014lf:Documents:lang-tech-papers:2017:LRTs-LATEX:img:LRT-Comp.png |
| Figure 1: |

1. **2. LITERATURE REVIEW: PREVIOUS WORK ON THE TOPIC**

LRT Placement test has become a trend used to measure language proficiency, it is already implemented for English, German, Spanish and others; meanwhile it is still not having real empirical for Arabic, Arabic language is an attractive challenge area for applying LRT tests.

Many others had deeply investigated this area, LRT or recognizing words started since 2000 as mentioned by Schmitt, on the other hand (Meara and Jones, 1987) is a previous example for words & non-words (LRT) testing, another English version called LexTALE were provided in 2012 by Lemhöfer and Broersma.

Another example was proposed by (Balota et al., 2007), they constructed a non-word database by manually changing one or more letters starting with known English words.

On the other side, ARC nonword database (Rastle et al., 2002) that contains monosyllabic non words which follow the phono tactic and orthographic rules of (Australian) English, The database only provides the non-words, but does not rank them according to their quality another approach is WordGen (Duyck et al., 2004), which is an interactive tool for generating non words.

Non word properties that are similar to the ones we use for ranking, e.g. neighborhood size, position specific bigram frequency etc Wuggy (Keuleers and Brysbaert, 2010) builds on

WordGen but introduces syllable template to build non words that more closely resemble a certain word.

A previous attempt for finding similar Arabic LRT could be listed as the following:

Baharudin et al. (2014) created an Arabic test of 40 words, all are right vocabularies but some not famous/frequently used, this exam called Test of Arabic Vocabulary (TAV).

In 2015 Test of Arabic Checklist founded by Ricks, a new Arabic vocabulary test was created, it consists of 40 words and 40 non words, format is similar to LexTALE method, dialects was ignored.

Till the moment, no adaptation for an Arabic LRT test, some challenges and obstacles ahead, Arabic considered a multi-formed language (one vocabulary could hold more than one meaning based on the gotten form through “diacritization/Tashkeel”), from this point of view importance of Diacritics for LRTs in a recent study, the researchers added a new parameter to Ricks’s test; They constructed a diacritized test, where they partially diacritized the test stimuli (words and nonwords) and applied a form of relaxation that drops some diacritics. It was shown that diacritics play a vital role in words recognition, especially for beginner and intermediate learners.

1. **3. THEORETICAL CONCERNS / SITUATING THE STUDY IN THE FIELD**

Meantime finding relevant reviewed literature, we can notice intersection between what we are going to do with what is already exists, many common points between current LRT for Arabic or other languages with our proposed diacritized Arabic LRT placement test; but we can distinguish our as it is diacritized and will be proved on site and conclude the differences against the previous ones and to show the progress.

1. **4. RESEARCH QUESTIONS AND THE WORKING HYPOTHESIS**

First Question came to mind if using language model could produce a good non words, another issue related to possibility of finding words of various fields?; the last thing related to finding a good algorithm that could do diacritization in a smart efficient way.

1. **5. METHODOLOGY**

Here, we will present the scenarios that describes the flow of data in the system and illustrates the main UML diagrams of the application. Finally we will describe the prototype of the model, main hypothesis supposed to be done are listed here:

1. Creating corpus, storing it in a repository data base.
2. Creating an application linked to the created repository and based an algorithm generating an exams and storing result in an XML files to ensure high performance when real exam session is initiated so the exam page will read a pre-defined XML exams presented in a well format harmonized with UX & HCI concepts.
3. Exam will be done in a few minutes and assess the used directly after submitting answer button.
4. Comparative study between our diacritized LRT approach against the not diacritized one.
5. **6. POSSIBLE OBSTACLES**

Building a comprehensive corpus that contains a diversity of vocabularies of different fields is a challenge task since a trusted standard modern language resources are wanted, such resources should contains sports, science, history, art, political and other fields vocabulary; Suggested resources are the main stream content of well-known news media as Aljazeera.

On the other hand, generating non words process has more than one way/algorithm, what should be done is to determine the best one that generate most powerful non words with least error rates; some algorithms using language model “nGram”, other using statistics probability model and others.

Finally, diacritizations process still represents a big challenge for generated non words, a new algorithm that will execute this job is suggested by me, other algorithms are already available, but we need to judge the better one to be used.

1. **7. RELEVANCE OF YOUR STUDY (contribution)**

Lexical recognition tests LRT has been become a trend to measure language proficiency for foreign, it had been applied for many languages as English, French, Spanish and German; mean time there are some attempts to apply LRT for Arabic but researches in this field still shallow and it is not exists as diacritized version, since Arabic vocabularies has polymorphism “multi-forms” of same vocabulary, it is important to add diacritization to reduce ambiguity and maturity; here is the proposed contribution is going to be by creation a new Arabic diacritized LRT (words and non-words) version.

1. **8. TENTATIVE THESIS CHAPTER OUTLINE**

Creating Arabic diacritized LRT version has many challenges of NLP tasks especially for Arabic language as it has its own flavor “polymorphism” and diaritization, here is the planned research sections that could be handled sequentially to obtain on Arabic LRT:

1. Words Data preparation: below operations are included to accomplish this task:
   1. Crawling
   2. Tokenization
   3. Word processing
   4. build a corpus
   5. Words diacritization
2. NonWords Generations:
   1. Using the selected algorithm to generate good nonWords.
   2. NonWords diacritization
3. Exam Evaluation
   1. Preparing the exam that includes words and nonWords vocabularies.
4. Questions Formulation
   1. Exam auto-correction
   2. Comparative study: to measure proficiency of this exam towards already existing Arabic LRT version (without diacritization) and with the real student level declared by the teacher.
5. **9. RESEARCH PLAN / TIMETABLE**

Here is a timeline schedule to complete proposed thesis with proof of concept:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Task Name** | **Duration** | **Start** | **Finish** | **Resource Property** |
| Software Development | 168 days | Mon 4/16/18 | Wed 12/5/18 |  |
| Scope | 12 days | Mon 4/16/18 | Tue 5/1/18 | Project Manager |
| Analysis/Software Requirements | 28 days | Wed 5/2/18 | Fri 6/8/18 | Analyst |
| Design | 21 days | Sat 6/9/18 | Fri 7/6/18 | Developer |
| Development | 42 days | Sat 7/7/18 | Mon 9/3/18 | Developer |
| Testing | 14 days | Tue 9/4/18 | Fri 9/21/18 | Testers |
| Training | 7 days | Sat 9/22/18 | Mon 10/1/18 | Trainers |
| Documentation | 14 days | Tue 10/2/18 | Fri 10/19/18 | Project Manager |
| Pilot | 14 days | Sat 10/20/18 | Wed 11/7/18 | Developer |
| Deployment | 7 days | Thu 11/8/18 | Fri 11/16/18 | Developer |
| Post Implementation Review | 7 days | Sat 11/17/18 | Mon 11/26/18 | Technical Communicators |
| Software development template complete | 7 days | Tue 11/27/18 | Wed 12/5/18 | Developer |

**10. BIBLIOGRAPHY**

[1] 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, Poznan, Poland. pp. 473–477. URL: <http://www.ltl.uni-due.de/wp-content/uploads/nonwords-ltc20151.pdf>

[2] Ricks, R., 2015. The Development of Frequency-Based Assessments of Vocabulary Breadth and Depth for L2 Arabic.

[3] Westbury, Chris, Geoff Hollis, and Cyrus Shaoul, 2007.

Lingua: the language-independent neighborhood generator of the University of Alberta, The Mental Lexicon, 2(2):271–284.

[4] Piotr Bojanowski∗ and Edouard Grave∗ and Armand Joulin and Tomas Mikolov, 2017. Enriching Word Vectors with Subword Information. URL: <https://arxiv.org/pdf/1607.04606.pdf>

[5] Hamed, O., Zesch, T., 2015. The Role of Diacritics in Designing Lexical Recognition Tests for Arabic, in: 3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5–6 November 2017, Dubai, United Arab Emirates. URL: <https://ac.els-cdn.com/S1877050917321579/1-s2.0-S1877050917321579-main.pdf?_tid=22299690-e8e3-11e7-a4f6-00aab0f6c&acdnat=1514144968_ebb71f0c1d693150f6b5fb334bd776e3>

[6] Schmitt, Norbert, 2000. Vocabulary in language teaching. Ernst Klett Sprachen.

[7] Meara, Paul and Glyn Jones, 1987. Tests of vocabulary size in English as a foreign language. Polyglot, 8(1):1–40.

[8] Lemhöfer, Kristin and Mirjam Broersma, 2012. Introducing lextale: A quick and valid lexical test for advanced learners of english. Behavior Research Methods, 44(2):325–343.

[9] Rastle, Kathleen, Jonathan Harrington, and Max Coltheart, 2002. 358,534 nonwords: The arc nonword database. The Quarterly Journal of Experimental Psychology:

Section A, 55(4):1339–1362.

[10] Keuleers, Emmanuel and Marc Brysbaert, 2010. Wuggy: A multilingual pseudoword generator. Behavior Research Methods, 42(3):627–633.

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