An Online Readability Leveled Arabic Thesaurus

Zhengyang Jiang, Nizar Habash, Muhamed Al Khalil
Computational Approaches to Modeling Language (CAMeL) Lab
New York University Abu Dhabi
`{zj522,nizar.habash,muhamed.alkhalil}@nyu.edu`

Abstract

This demo paper introduces the online Readability Leveled Arabic Thesaurus interface. For a given user input word, this interface provides the word's possible lemmas, roots, English glosses, related Arabic words and phrases, and readability on a five-level readability scale. This interface builds on and connects multiple existing Arabic resources and processing tools. This one-of-a-kind system enables Arabic speakers and learners to benefit from advances in Arabic computational linguistics technologies. Feedback from users of the system will help the developers to identify lexical coverage gaps and errors.

1 Introduction

Arabic is one of the six UN official languages, the language of millions of people in the Arab world, as well as the liturgy language of Muslims. It is also a language known for its difficulty for new learners; and its Modern Standard Arabic (MSA) form used in education and the media is technically not the native form spoken by modern-day Arabs, who speak a variety of its dialects. As such, there is a great need to have user-friendly interfaces for searching on Arabic words and their relations targeting Arabic teachers and learners. However, a small minority of dictionaries in general (and none in Arabic to our knowledge) specify the readability level of their words, let alone their lexical relations with other words.

The system we present in this paper exploits a number of developments in Arabic natural language processing (NLP) by different groups of researchers (Black et al., 2006; Graff et al., 2009; Taji et al., 2018; Obeid et al., 2020; Al Khalil et al., 2020) to develop a new online thesaurus that (a) supports different search modes (inflected word, lemma, root and English gloss), (b) provides five types of lexical relations (synonyms, antonyms, hypernyms, hyponyms and related), and (c) indicates the readability level of the word on a five-scale system. This interface allows Arabic speakers and learners to benefit from advances in Arabic NLP technologies. And by exposing these technologies to a large number of users, we expect their feedback will help the researchers who developed the computational and lexical components to identify gaps and errors.

A live link to the demo is available at: `http://samer.camel-lab.com/`.

2 Background and Related Work

In this section, we present the main challenges for processing Arabic, and the various databases and NLP tools we use in developing our system.

Arabic Linguistic Considerations Arabic is a morphologically rich and orthographically ambiguous language. Words have many inflected forms varying in terms of gender, number, person, case, aspect, mood, voice, as well as a large number of attachable clitics, such as pronominal objects and prepositions. Short vowels and consonant doubling are indicated using optional diacritical marks that are mostly elided resulting in a high degree of ambiguity. For example the word فردها *frdhA* has four core lemmas (or

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: `http://creativecommons.org/licenses/by/4.0/`.

Proceedings of the 27th International Conference on Computational Linguistics, pages 59–63
Barcelona, Spain (Online), December 12, 2020.

lexical abstractions over all inflections): the verbs فَرَّد *far∼ad* ‘individualize, separate in units’, and رَدّ *rad∼* ‘answer, return’; and the nouns فَرْد *fard* ‘individual, unit’ and رَدّ *rad∼* ‘response, return’.[1] The root of the first and third lemmas is [ف.ر.د] *[f.r.d]*, and the root of the other lemmas is [ر.د.د] *[r.d.d]*.

Arabic Natural Language Processing Tools To address the morphological complexity and orthographic ambiguity, we make use of an open source toolkit for Arabic NLP, Camel Tools (Obeid et al., 2020). We embedded its morphological analyser, CALIMA Star (Taji et al., 2018), in our online system to determine all the lemmas associated with a user’s input in word mode. CALIMA Star extends the Standard Arabic Morphological Analyzer (SAMA) (Graff et al., 2009) with a number of additional morphological features. The database covers over 40 thousand lemmas and links each to a part-of-speech (POS), root and English gloss, all of which can be searched on in our interface.

Lexical Modeling with Arabic WordNet Arabic WordNet (AWN) is a public lexical database of semantic relations between words in Arabic (ElKateb et al., 2006; Black et al., 2006). AWN was developed based on the methods used in EuroWordNet (EWN) (Rodríguez et al., 1998), and is directly mappable to it and to the Princeton Wordnet of English (Fellbaum, 1998). The AWN database currently has 16,066 entries, out of which 5,036 are multi-word phrases. Basic entries are represented as lemmas abstracted from morphological inflections. The AWN entries are organized in 14,284 synonym sets (synsets), some of which are connected through lexical relations such as antonymy, hypernymmy, and others.

The AWN is the *Thesaurus* backbone of our system. We focus on five types of lexical relations: synonym, antonyms, hyponyms, hypernyms and *related*. The first four relations are defined as in AWN. Most other relevant AWN relations are mapped to the *related* relation. In our setup, the synonyms of a lemma *x* are the union of all the other lemmas in all the synsets containing the lemma *x*. Similarly, the antonyms of a lemma *x* are the union of all the lemmas in the synsets that are in an antonym relation with the synsets containing the lemma *x*.

Arabic Readability Modeling readability levels is relevant to a range of NLP tasks from developing language education applications to user profiling. Much work has been done on readability leveling and its assessment and specification in English leading to the development of many resources and tools. However, this is not the case for many other languages. A recently developed Arabic Readability Lexicon have filled this gap for Arabic (Al Khalil et al., 2020). Language professionals manually annotated a 26,578-lemma lexicon with a five-level readability scale for MSA targeting native speakers. The levels are as follows: Level I (Grade 1, age 6), Level II (Grade 2-3, age 7-8), Level III (Grade 4-5, age 9-10), Level IV (Grade 6-8, age 11-14), Level V (specialist, age 15 and above).

3 Design and Implementation

We start with a discussion of the design desiderata of our system, followed by details of the database preparation, back-end functionalities and front-end interface.

3.1 Design Specifications

We designed our interface with the following considerations in mind.

- **Handling Arabic Ambiguity and Rich Morphology** The system needs to provide the ability to search on an inflected Arabic word form by relating it to its lemma and POS.
- **Multiple Search Modes** In addition to searching on an inflected word, the system needs to provide the ability to directly search on a lemma with its POS, on an Arabic root, and on an English gloss. Multiword phrases should be also searchable using component words.
- **Rich Lexical and Readability Information** In addition to the readability level of the lemmas associated with the search mode, the system needs to provide lexically related words and phrases.
- **Friendly Navigation** The system needs to allow hyperlink search: the user can click on underlined components of returned search results to explore the network of the used databases.

[1] Arabic transliteration is presented in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007).

#	Key	Value	Example
1	lemma#pos	English	سَفَر#noun -> journey, travel, trip
2	lemma#pos	root	كِتاب#noun ‘book’ -> [ك ت ب]
3	root	(lemma#pos)	[ك ت ب] -> كِتاب#noun ‘book’,كاتِب#noun ‘author’
4	English	(lemma#pos)	book -> كِتاب#noun, سِفْر#noun
5	lemma#pos	readability	كِتاب#noun ‘book’ -> Level 1
6	lemma#pos	synonyms (lemma#pos)	كِتاب#noun ‘book’ -> سِفْر#noun ‘book’
7	lemma#pos	antonyms (lemma#pos)	اِمْرَأَة#noun ‘woman ->رَجُل#noun ‘man’
8	lemma#pos	hypernyms (lemma#pos)	أَمِير#noun ‘prince’ -> إِنْسان#noun ‘human’
9	lemma#pos	hyponyms (lemma#pos)	إِنْسان#noun ‘human’ -> أَمِير#noun ‘prince’
10	lemma#pos	related expressions (lemma#pos)	أُمّ#noun ‘mother’ -> دَلَّل#verb ‘pamper’
11	lemma#pos	matching phrases (lemma#pos)	بَيْت#noun ‘house’ -> بيت الشباب#noun ‘hostel’

Figure 1: Contents of the Look-up Tables (examples are not complete entries)

3.2 Database Preparation

Our system uses 11 look-up tables to provide the needed support for all the search modes described above. The look-up table key and value information are listed in Figure 1. The primary search key across most of the look-up tables is the lemma and POS concatenated by the hash sign #.

Morphological Look-up Tables Tables 1 through 4 link the lemma#pos to the English gloss and root, and the English gloss and root to lemma#pos. These look-up tables are populated from the Camel Tools databases. The first two provide the specific root and English gloss of a lemma#pos; while the last two link the root and English gloss to *all* the lemma#pos values sharing the same root, or English gloss, respectively. We lower case English words when using them as look-up keys.

Readability Look-up Table Table 5 links the lemma#pos to the readability level. It is populated from the Arabic Readability Lexicon.

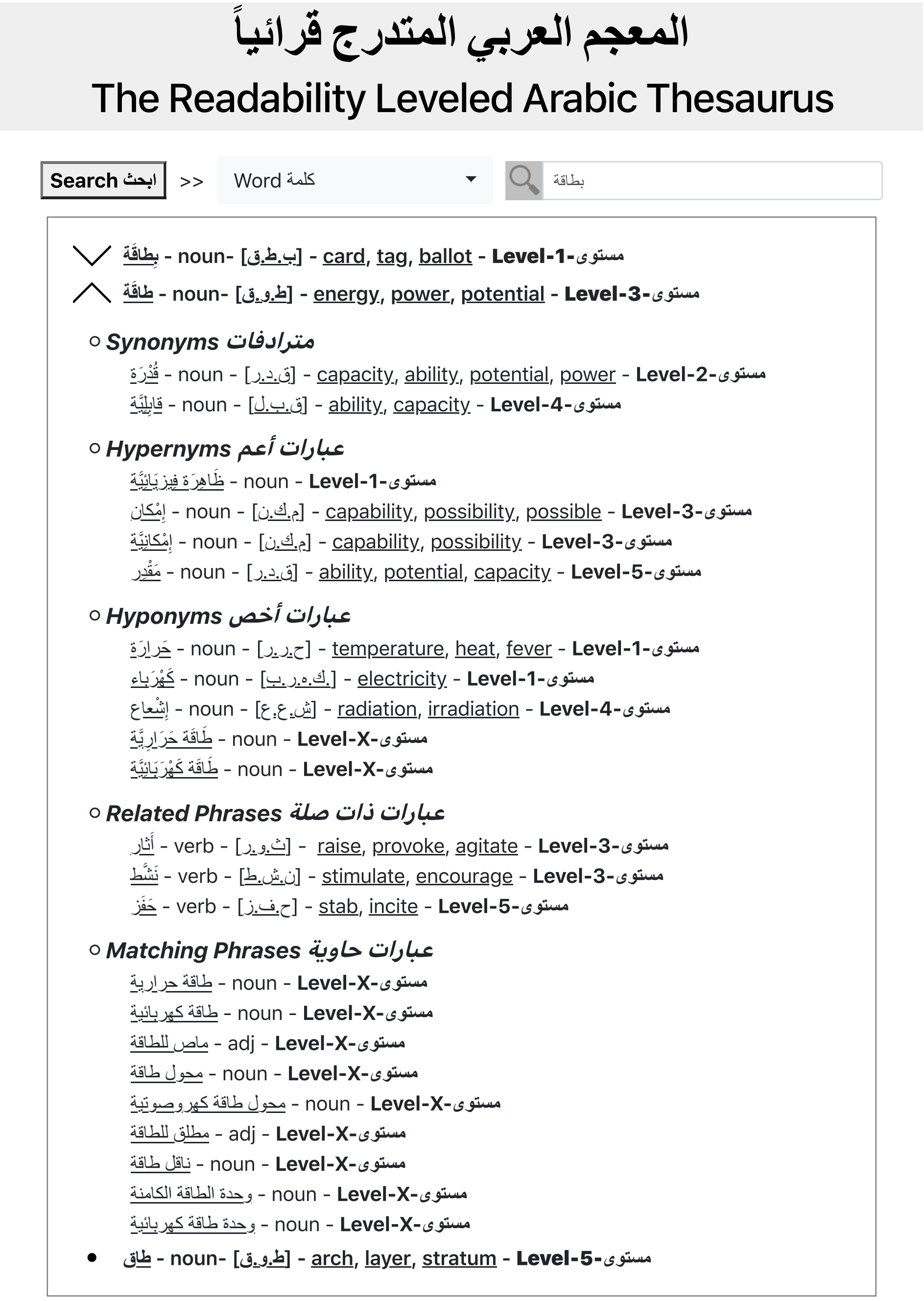
Semantic Relation Look-up Tables Tables 6 through 11 link a key lemma#pos to a list of other lemma#pos entries that are in specific semantic relationship to the key lemma#pos. The relation types include synonyms, antonyms, hypernyms, hyponyms, related expressions, and matching phrases (exact lemma#pos match within a larger multiword phrase).

The semantic relation look-up tables are populated from AWN after matching AWN lemmas with Camel Tools lemmas. One challenge we faced is that AWN lemmas and POS are defined slightly differently from Camel Tools, e.g., the lemma for the word ‘aroma’ is شَذَا#n *šaðaA*#n in AWN but شَذاً#noun *šaðAã*#noun in Camel Tools. We match the lemmas in an offline process using the MADAMIRA Arabic morphological analyzer and disambiguator (Pasha et al., 2014), which uses the same lemmas and POS as Camel Tools. We run the AWN word and multiword entries through MADAMIRA, and then select the best matches using Levenshtein edit distance, and manual mapping rules between the POS categories.

3.3 Implementation

Our back-end was implemented in Python using Flask.[2] We embedded Camel Tools' Arabic morphological analyzer to process user input Arabic words, and map them to lists of ambiguous lemma#pos entries. These entries are used as keys to look up other features and related lemma#pos entries in the look-up tables. The display of the information includes dynamically creating hyperlinks to search on the lemma#pos, root and English gloss. For the front-end, we used simple JavaScript to control the look and feel of the interface.

[2] https://flask.palletsprojects.com/en/1.1.x/



المعجم العربي المتدرج قرائياً

The Readability Leveled Arabic Thesaurus

بِطاقَة - noun- [ب.ط.ق] - card, tag, ballot - **Level-1-مستوى**

طاقَة - noun- [ط.و.ق] - energy, power, potential - **Level-3-مستوى**

○ ***Synonyms مترادفات***

قُدْرَة - noun - [ق.د.ر] - capacity, ability, potential, power - **Level-2-مستوى**

قابِلِيَّة - noun - [ق.ب.ل] - ability, capacity - **Level-4-مستوى**

○ ***Hypernyms عبارات أعم***

ظَاهِرَة فِيزِيَائِيَّة - noun - **Level-1-مستوى**

إِمْكان - noun - [م.ك.ن] - capability, possibility, possible - **Level-3-مستوى**

إِمْكانِيَّة - noun - [م.ك.ن] - capability, possibility - **Level-3-مستوى**

مَقْدِر - noun - [ق.د.ر] - ability, potential, capacity - **Level-5-مستوى**

○ ***Hyponyms عبارات أخص***

حَرارَة - noun - [ح.ر.ر] - temperature, heat, fever - **Level-1-مستوى**

كَهْرَباء - noun - [ك.ه.ر.ب.] - electricity - **Level-1-مستوى**

إِشْعاع - noun - [ش.ع.ع] - radiation, irradiation - **Level-4-مستوى**

طَاقَة حَرَارِيَّة - noun - **Level-X-مستوى**

طَاقَة كَهْرَبَائِيَّة - noun - **Level-X-مستوى**

○ ***Related Phrases عبارات ذات صلة***

أَثار - verb - [ث.و.ر] - raise, provoke, agitate - **Level-3-مستوى**

نَشَّط - verb - [ن.ش.ط] - stimulate, encourage - **Level-3-مستوى**

حَفَز - verb - [ح.ف.ز] - stab, incite - **Level-5-مستوى**

○ ***Matching Phrases عبارات حاوية***

طاقة حرارية - noun - **Level-X-مستوى**

طاقة كهربائية - noun - **Level-X-مستوى**

ماص للطاقة - adj - **Level-X-مستوى**

محول طاقة - noun - **Level-X-مستوى**

محول طاقة كهروصوتية - noun - **Level-X-مستوى**

مطلق للطاقة - adj - **Level-X-مستوى**

ناقل طاقة - noun - **Level-X-مستوى**

وحدة الطاقة الكامنة - noun - **Level-X-مستوى**

وحدة طاقة كهربائية - noun - **Level-X-مستوى**

- طاق - noun- [ط.و.ق] - arch, layer, stratum - **Level-5-مستوى**

Figure 2: Search Results for the word بطاقة *bTAqħ*

3.4 Example

We illustrate the functionality of our system with the example in Figure 2. For the undiacritized Arabic word بطاقة *bTAqħ* input by a user in Word mode, our server processes the input with the Camel Tools' morphological analyzer generating three analyses: بِطاقَة#noun *biTaAqaħ* 'ballot', طاقَة#noun *TaAqaħ* 'energy', and طاق#noun *TaAq* 'arch'. Their English gloss, readability and root details are obtained from look-up tables 1-5 (in Figure 1). These three analyses are shown as the first level result. Once the user clicks on the arrow sign on the side of a first level result, the second level results, or the relations associated with the first level result, are displayed. Those relations' keys (lemma#pos) are obtained from look-up tables 6-11. For each key in those relations, we search look-up tables 1-5 to obtain their English gloss, readability and root information.

4 Conclusion and Future Work

We presented an online readability leveled Arabic thesaurus, created through integrating a number of Arabic NLP tools and data sets. We plan to continue improving on the various elements used in building this interface. The most immediate use of the interface is as part of an effort to simplify Arabic modern novels to younger readers (Al Khalil et al., 2017).

Acknowledgments

This project was supported by a New York University Abu Dhabi Research Enhancement Fund. We would like to thank Salam Khalifa and Ossama Obeid for helpful conversations, and interface design and implementation advice. We thank the anonymous reviewers for their feedback and comments.

References

Muhamed Al Khalil, Nizar Habash, and Hind Saddiki. 2017. Simplification of Arabic masterpieces for extensive reading: A project overview. *Procedia Computer Science*, 117:192–198.

Muhamed Al Khalil, Nizar Habash, and Zhengyang Jiang. 2020. A large-scale leveled readability lexicon for standard Arabic. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3053–3062, Marseille, France, May. European Language Resources Association.

William Black, Sabri Elkateb, Horacio Rodríguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Introducing the Arabic WordNet project. In *Proceedings of the third international WordNet conference*, pages 295–300. Citeseer.

Sabry ElKateb, William Black, Horacio Rodríguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Building a WordNet for Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Genoa, Italy.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press. http://www.cogsci.princeton.edu/∼wn [2000, September 7].

David Graff, Mohamed Maamouri, Basma Bouziri, Sondos Krouna, Seth Kulick, and Tim Buckwalter. 2009. Standard arabic morphological analyzer (sama) version 3.1. *Linguistic Data Consortium LDC2009E73*.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France, May. European Language Resources Association.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T 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*, volume 14, pages 1094–1101.

Horacio Rodríguez, Salvador Roca, Piek Vossen, Laura Bloksma, Wim Peters, Antonietta Alonge, Francesca Bertagna, and Adriana Roventini. 1998. The top-down strategy for building eurowordnet: Vocabulary coverage, base concepts and top ontology. *Computers and the Humanities*, 32:117–152, 03.

Dima Taji, Salam Khalifa, Ossama Obeid, Fadhl Eryani, and Nizar Habash. 2018. An Arabic morphological analyzer and generator with copious features. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 140–150, Brussels, Belgium, October. Association for Computational Linguistics.