An Arabic-Multilingual Database with a Lexicographic Search Engine

Mustafa Jarrar(✉) and Hamzeh Amayreh

Birzeit University, Birzeit, Palestine
mjarrar@birzeit.edu, hamayreh@staff.birzeit.edu

Abstract. We present a lexicographic search engine built on top of the largest Arabic multilingual database, allowing people to search and retrieve translations, synonyms, definitions, and more. The database currently contains about 150 Arabic multilingual lexicons that we have been digitizing, restructuring, and normalizing over 9 years. It comprises most types of lexical resources, such as modern and classical lexicons, thesauri, glossaries, lexicographic datasets, and (bi/)tri-lingual dictionaries. This is in addition to the Arabic Ontology – an Arabic WordNet with ontologically cleaned content, which is being used to reference and interlink lexical concepts. The search engine was developed with the state-of-the-art design features and according to the W3C's recommendation and best practices for publishing data on the web, as well as the W3C's Lemon RDF model. The search engine is publicly available at (https://ontology.birzeit.edu).

Keywords: Arabic · Multilingual lexicons · Online dictionary · Language resources · Lexical semantics · Lexicographic search · W3C lemon · RDF · NLP

1 Introduction and Motivation

The increasing demands to use and reuse dictionaries (of all types) in modern applications have shifted the field of lexicography to be a multidisciplinary domain, engaging ontology engineering [18, 19, 22], computational linguistics [1, 17, 25], and knowledge management [12, 16, 23, 24]. Dictionaries are no more limited to hard copies and are not only used by humans; they are becoming important for IT applications that require natural language processing [6–8, 20], in addition to the need to access them electronically. In response to these demands, there have been several efforts to digitize, represent, and publish them online. As will be discussed later, the ISO37 has released more than 50 standards in the past 15 years related to terminology and lexical resources, in addition to several W3C recommendations e.g., SKOS, Lemon [22], and the Linguistic Linked Open Data Cloud [7].

Although there are many lexicons available on the internet for most languages, especially English, few Arabic lexicons are available in digital forms [1, 5, 25]. This lack of such digital resources has limited the progress in Arabic NLP research [21], and has also led many people to use statistical machine translation tools (e.g., Google Translate) in place of dictionaries [4].

E. Métais et al. (Eds.): NLDB 2019, LNCS 11608, pp. 234–246, 2019.
https://doi.org/10.1007/978-3-030-23281-8_19

In this paper, we present the digitization of 150 Arabic multilingual lexicons and a lexicographic search engine built on top of them, covering many domains such as, natural sciences, technology, engineering, health, economy, art, humanities, philosophy, and more. The digitization process was carried out over 9 years, as most lexicons had to be manually typed, then restructured and normalized. The copyright owners of all lexicons were contacted individually for a permission to digitize and use their lexicons. To the best of our knowledge, our database is currently the largest Arabic lexicographic database, compromising about 2.4 million multilingual lexical entries and about 1.1 million lexical concepts. The search engine was designed according to W3C's recommendations, especially the Lemon model which is important for referencing and linguistic data linking purposes. The ranking strategy used in the search engine is a combination metric of lexicon-renown and concept-citation.

The rest of this paper proceeds as follows: In Sect. 2, we overview related work. We elaborate on the construction of the lexicographic database in Sect. 3. Section 4 presents the search engine, its architecture, URLs design, ranking strategies, and usability. Finally, we conclude and discuss future work in Sect. 5.

2 Related Work

We first present recent standards for representing and publishing linguistic data, then we overview related digital lexicographic resources and repositories.

In response to the increasing demands to use and reuse linguistic resources in modern applications, there have been many efforts to standardize the way linguistic resources are structured, represented, and published on the web. The ISO37 produced over 50 standards in the recent years in this direction. For example, the ISO24613 is a lexical markup framework (LMF) to represent lexicons in a machine-readable format; the ISO860 is concerned with the harmonization of concepts, concept systems, definitions and terms; the ISO16642 supports the development, use, and exchange of terminological data between different IT applications. W3C has also developed several recommendations related to linguistic data sources. For example, SKOS provides a way to represent thesauri, classification schemes, subject headings, and taxonomies within the framework of the Semantic Web. The W3C's Lemon RDF model [22] aims at enabling lexicons to be used by ontologies and NLP applications. It can be used to describe the properties of lexical entries and their syntactic behavior, encouraging reuse of existing linguistic data. The importance of Lemon is that, it was developed based on the W3C recommendations for Open Linked Data [2]. The Linguistic Linked Open Data Cloud (LLOD) [7] was initiated as a collaborative effort to interlink the lexical entries of different linguistic resources using Lemon.

A new ambitious project, called PanLex, aims at building the world's largest lexical database [3], with 2500 dictionaries for 5700 languages. Its objective is to be a bilingual translation-oriented database, offering about 1.3 billion translation pairs. Compared with our work, PanLex offers only bilingual translations, rather than a lexicographic database with definitions, synonyms, and other lexicographic features, and it does not support a large number of Arabic lexicons.

There have also been other related initiatives aiming to integrate wordnets with other resources. BabelNet [24] is a multilingual encyclopedic dictionary covering 284 languages, as an integration of many wordnets, Wiktionary, Wikipedia, GeoNames, and more. BabelNet is a semantic network connecting concepts and named entities. Similarly, ConceptNet 5.5 [23] is an open multilingual knowledge graph connecting words and phrases with labeled edges. It links the Open Mind Common Sense with Multi Wordnet, Wiktionary, Wikipedia, and OpenCyc. Compared with our work, both, BabelNet and ConceptNet, aim at building encyclopedic knowledge graphs, and their linguistic information is limited to wordnets rather than targeting a large number of dictionaries as we try to do. Additionally, their support of Arabic is limited to Arabic Wikipedia and Arabic WordNet that is quite small (only 11 k synsets).

The number of available structured Arabic lexicons in digital format is indeed limited [1, 21, 25]. Earlier attempts to represent Arabic morphological lexicons in ISO LMF standard can be found in [25], and in [6] to represent Dutch bilingual lexicons, including Arabic. A lexicon called Al-Madar [1] was developed and represented using the ISO LMF standard. Similarly, Al-Qamus Almuhit was digitized and represented in the ISO LMF and later in the Lemon model [20]. A preliminary progress in digitizing several Hadith lexicons is reported in [21]. Nevertheless, none of the lexicons above is accessible online.

There are several online portals offering lexicographic search (e.g., almaany.com, lisaan.net, almougem.com, albaheth.info, ejtaal.net, alburaq.net), each comprises only a small number of lexicons. More importantly, most of the content in these portals is partially structured (i.e. available in flat text), as they allow people to search for a word, and the paragraphs that include this word as a headword will be retrieved.

It is worth noting that, modern Arabic lexicons are mostly the production of two authoritative institutions, the ALECSO that produced about 50 lexicons, and the Arabic Academy in Cairo that produced about 20 lexicons. The majority of these lexicons were digitized and included in our database. Additionally, SAMA[1] and Sarf[2] are two Arabic morphological databases that were designed for morphological analysis only. Both are being used to map between, and enrich, the lexical entries in our database.

3 Constructing the Lexicographic Database

This section presents our database, which contains about 150 lexicons that we have been digitizing from scratch. Although we were able to obtain some lexicons in digital flat text format, we had to type most lexicons manually. First, we tried to use OCR tools, but we failed due to their low quality. We also failed in crowdsourcing the digitization process among 300 students as most of them were uncareful. Afterwards, we contracted some careful students to type lexicons in MS Word format, and then gave the output to two experts to manually compare with original copies. The output was then converted into a preliminary ad hoc DB table. A lexicon that uses explicit and

[1] Developed by LDC, accessible at: https://catalog.ldc.upenn.edu/LDC2010L01.

[2] An open source project, accessible at: https://sourceforge.net/projects/sarf/.

steady markers (e.g. tab, comma, semicolon) to separate between different features was parsed and converted automatically; otherwise, such markers were manually added before parsing.

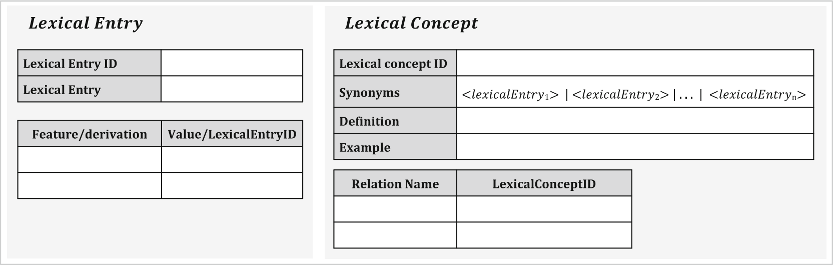
Lexicon Restructuring: the content of each lexicon was restructured separately, which was a semi-automated task (fully presented in [5]). Before overviewing this task, we first present a classification of our lexicons, and describe their internal structure and type of content:

- **Glossary**: a domain-specific lexicon, where each lexical entry is defined in a few lines. Advanced glossaries provide also synonyms, multilingual translation(s), and sometimes references to related lexical entries, e.g. similar, equivalent, or related.
- **Thesaurus**: sets of synonymous lexical entries. Each set might be lexicalized in one or more languages. A set might be also labeled with a part-of-speech tag.
- **Dictionary**: a list of lexical entries, each with some bi/trilingual translations.
- **Linguistic Lexicon**: a set of lexical entries, each with its linguistic features and sense(s). A lexical entry may have several meanings, which some lexicons designate into separate senses, while others combine them in one description. Lexicons may also provide linguistic features for each entry, e.g. root, POS, and inflections.
- **Semantic-variations lexicon**: a set of pairs of semantically close lexical entries and the differences between their meanings, (e.g. like $\sim$ love, pain $\sim$ ache).

To structure the content of such types of lexicons, we developed two general templates (see Fig. 1), where each lexicon was parsed and mapped to them, as the following:

1. Each lexical entry in every language, whether provided as a headword, a synonym, or a translation, was extracted and given a unique ID. The features of the lexical entry, (e.g. POS, lexical forms, inflections), were all extracted and stored in the Lexical Entry template, in an RDF-like format. Deciding whether two Arabic lexical entries of the same letters are the same is challenging, as they might be partially or non-diacritized [11].
2. Each meaning of every headword in every lexicon is considered a lexical concept, and is given an ID. This is straightforward in case of glossaries as each headword typically has only one meaning. Each set of synonymous entries in a thesaurus, and similarly each group of translations in a dictionary, is mapped into a lexical concept and is given an ID. However, in case of linguistic lexicons, the different senses of a lexical entry were each extracted, and mapped into a separate lexical concept. In case of references to other lexical concepts, (e.g. indicating semantic relations like related and similar), these relations were also extracted and stored in the Lexical Concept template. In this way, the Lexical Concept template was filled in, providing the concept ID, and if available its set of synonyms, definitions, examples, and relations.

Cleaning and Normalization: The content of these general templates was then cleaned and normalized before storing them in a relational database. As lexicons are typically designed to be printed and used as hard copies, new challenges are faced



Lexical Entry

Lexical Entry ID	
Lexical Entry	

Feature/derivation	Value/LexicalEntryID

Lexical Concept

Lexical concept ID	
Synonyms	<$lexicalEntry_1$> | <$lexicalEntry_2$> |… | <$lexicalEntry_n$>
Definition	
Example	

Relation Name	LexicalConceptID

Fig. 1. Lexical concept and lexical entry templates.

when converting them into a machine processable format. In what follows, we summarize some of these challenges – see [5] for more issues and details.

- *Challenges induced by ordering:* to maintain a proper alphabetical ordering in hard copies, many lexicons tend to re-arrange words in the lexical entry, such as: "accelerator (linear…)", "affinity (chemical)", "drawing (final)", "earth (the)", and "crush (to)". Detecting such cases and deciding whether to move the text between parentheses to the beginning or to keep the order intact is difficult. This is because parenthesis might be also used for other purposes, as in (e.g. "tube (pipe)", "academy (of art)"), which indicate synonymy and context, respectively. There are no markers that would help detecting and normalizing such lexical entries.
- *Subterm synonymy*: most lexicons use commas or other symbols to separate between synonyms (e.g. "benzene, benzol", "tie, bind"). Though it is easy to split them, we found many cases where the comma is used differently, e.g. to indicate a more specific meaning, as in "calomel electrode, calomel", "kelvin's scale, kelvin's absolute scale", and "liquid drier, drier". That is, if a term is synonymous with another term, and one is part of the other, it is likely to be a mistake or to indicate another more specific meaning. Such cases need to be manually reviewed and decided upon.
- *Long multiword lexical entries:* there are cases where a lexical entry is composed of many words, such as "buildings or other structures recurrent taxes on land". Such cases of long and "poor" lexical entries need to be manually reviewed. i.e. by considering it a definition, or excluding it.
- *Special characters*: The use of special characters in a lexical entry (e.g. quotations, punctuation marks, and brackets) is allowed if they were used intentionally as part of the lexical entry. Nevertheless, they are often introduced in lexicons as annotations. Therefore, they have to be filtered out and individually reviewed.
- *Character set*: Same characters and symbols have different encodings across different languages (e.g., the dash, quotations, punctuations, and whitespaces), which is not a problem in case of printed lexicons, but they are obstacles when digitizing lexicons. This issue is trickier in Arabic as there are also different versions of character sets and there are characters in Arabic which have the same orthography but with different encodings that need to be changed to use the same encoding version.

Addressing such challenges in a fully automatic manner is difficult. Therefore, we have developed a parsing framework, presented in [5], that first detect and filter out each individual issue (e.g., whether a lexical entry includes parenthesis, commas, subterm synonymy, long multiword entries, character set issue, etc.). The parsers then assign a category to each of these issues to indicate its nature. The output of the parsers includes also a suggested treatment, depending on the nature of the issue. Each category was then given to a linguist to review and confirm the suggested treatments. After normalizing lexical entries and features, the data was stored in a MySQL database and indexed for searching purposes. Table 1 illustrates some statistics about our database.

Table 1. Statistics of the Lexicographic Database – being extended.

Category	Lexical Concepts	Lexical entries	Synsets	Translations pairs	Glosses	Semantic relations
Total (Milions)	1.1 M	2.4 M	1.8 M	1.5 M	0.7 M	0.5 M
Sub Counts		1,100 K Arabic 1,100 K English 200 K French 3 K Others 1,300 K Single-word 1,000 K Multi-word	800 K Arabic 800 K English 200 K French 50 K Others	1,000 K English-Arabic 300 K English-French 200 K French-Arabic	400 K Arabic 300 K English 1 K Others	170 K Sub-super links 29 K Part-of links 260 K Has-Domain links 30 K Other links

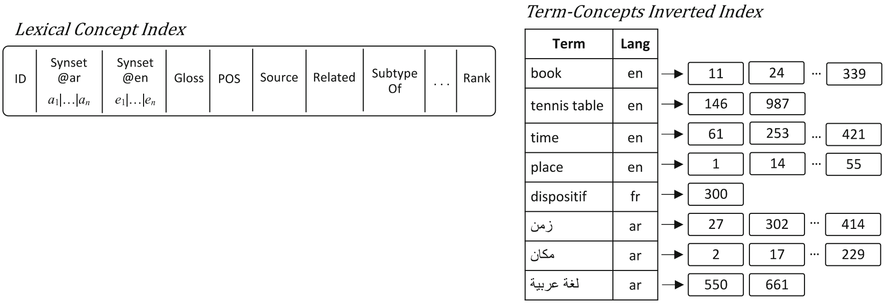
Copyrights: We contacted all lexicons' owners individually to get an official permission to digitize and include their lexicons in the search engine, a process that took several years, and although some refused, most of them accepted. Their main motivation was that the search engine displays the copyright symbol and the lexicon's name below each retrieved result, keeping their rights reserved. Additionally, when the lexicon's name is clicked (see Fig. 3), it shows the author(s), publisher, and links to their websites and to bookstores to purchase their lexicon.

Referencing lexical concepts in the Arabic Ontology

The Arabic Ontology is part of the database and is accessible in the search engine[3]. It can be seen and used as a formal Arabic wordnet built on the basis of a carefully designed ontology [9, 15]. It consists currently of about 1.3 K concepts that are also mapped to WordNet, BFO, and DOLCE, in addition to 11 K concepts that are being validated and mapped. The Arabic ontology is currently being used to reference lexical concepts in all lexicons; such that, each lexical concept is mapped (e.g., equal, or subtype) to a concept in the ontology. In this way, lexical concepts across all lexicons would be semantically linked; and since the ontology is mapped with other resources, it implies that lexical concepts would also be mapped to these resources. Presenting these mappings is beyond the scope of this article.

[3] http://ontology.birzeit.edu/concept/293198.

Data Indexing: To prepare our database for efficient search, we built two indexes, depicted in Fig. 2. The *Lexical Concept Index* aggregates relevant information for each concept in one record. It includes the concept ID, Arabic and English synsets, gloss (i.e. definition), semantic relations, and other features that need to be retrieved by the search engine. The computed rank for each lexical concept (as we will discuss later) is also pre-calculated and stored in this *Index*. The *Term-Concepts Inverted Index* is an inverted index that links between lexical entries and their lexical concepts. This inverted index was built by first collecting all lexical entries from all synsets, which can be single or multiple words (we call it *Term*). Second, by linking each of these terms with its posts (i.e. lexical concepts). This index was implemented using MySQL's full-text index, especially that it supports the generation of concordances. For search effectiveness, each of the terms in the inverted index was normalized and stemmed. Currently, *Lexical Concept* contains about 2.2 million records, whereas the inverted index contains about 1.1 million records, each having 25 postings on average.



Lexical Concept Index

ID	Synset @ar $a_1 \vert \ldots \vert a_n$	Synset @en $e_1 \vert \ldots \vert e_n$	Gloss	POS	Source	Related	Subtype Of	. . .	Rank

Term-Concepts Inverted Index

Term	Lang				
book	en	→ 11	24	···	339
tennis table	en	→ 146	987		
time	en	→ 61	253	···	421
place	en	→ 1	14	···	55
dispositif	fr	→ 300			
زمن	ar	→ 27	302	···	414
مكان	ar	→ 2	17	···	229
لغة عربية	ar	→ 550	661		

Fig. 2. Main Indexes.

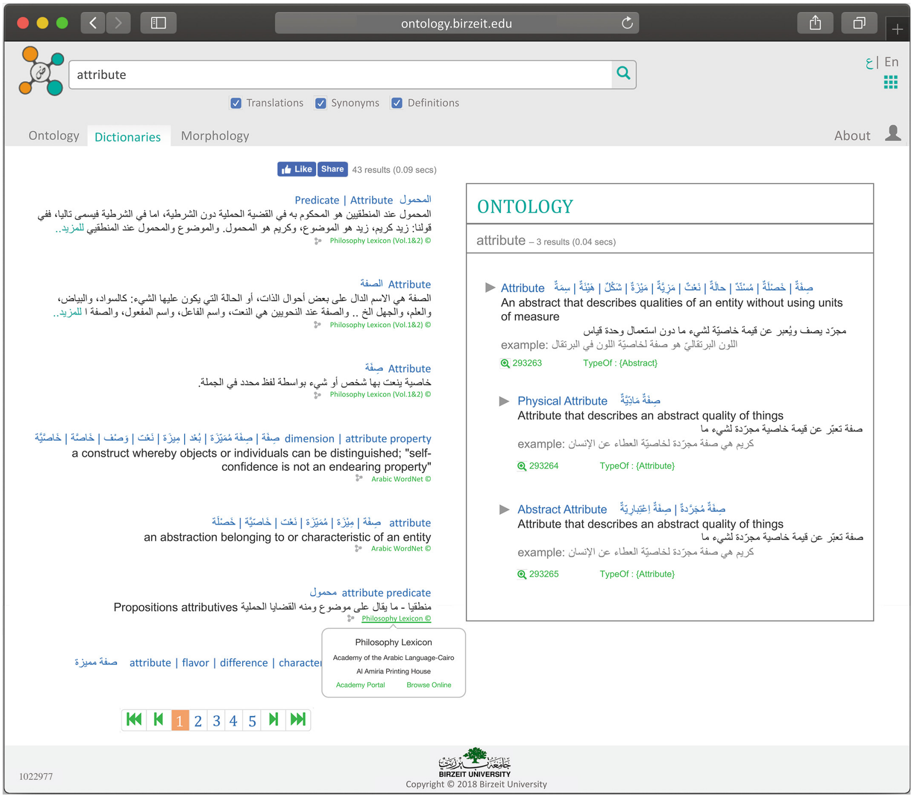
4 Building a Lexicographic Search Engine

Figure 3 illustrates a screenshot of the search engine. It allows people to search for translations, synonyms and definitions from the 150 lexicons and the Arabic Ontology, and filter the results. The engine is designed based on a set of RESTful web services (Fig. 4), which query the database and return the results in JSON format that is then rendered at our front-end, and can be also used by third-party applications.

4.1 URLs Design

The URLs in the search engine are designed according to the W3C's Best Practices for Publishing Linked Data [2], including the Cool URIs, simplicity, stability, and linking best practices, as described in the following URL schemes. This allows one to also explore the whole database like exploring a graph:

- *Term*: Each term (i.e., affix, word, or multiword expression) is given a URL: `http://{domain}/term/{term}`, which retrieves the set of all lexical concepts, in all lexicons, that are lexicalized using this term, i.e. that have this term as



43 results (0.09 secs)

Predicate | Attribute المحمول
المحمول عند المنطقيين هو المحكوم به في القضية الحملية دون الشرطية، اما في الشرطية فيسمى تاليا، ففي قولنا: زيد كريم، زيد هو الموضوع، وكريم هو المحمول. والموضوع والمحمول عند المنطقيي للمزيد..
Philosophy Lexicon (Vol.1&2) ©

Attribute الصفة
الصفة هي الاسم الدال على بعض أحوال الذات، أو الحالة التي يكون عليها الشيء: كالسواد، والبياض، والعلم، والجهل الخ .. والصفة عند النحويين هي النعت، واسم الفاعل، واسم المفعول، والصفة ا للمزيد..
Philosophy Lexicon (Vol.1&2) ©

Attribute صِفَة
خاصية ينعت بها شخص أو شيء بواسطة لفظ محدد في الجملة.
Philosophy Lexicon (Vol.1&2) ©

attribute property | dimension صِفَة | صِفَة مُمَيِّزَة | بُعْد | مِيزَة | نَعْت | وَصْف | خَاصَّة | خَاصِّيَّة
a construct whereby objects or individuals can be distinguished; "self-confidence is not an endearing property"
Arabic WordNet ©

attribute صِفَة | مِيْزَة | مُمَيِّزَة | نَعْت | خَاصِّيَّة | خَصْلَة
an abstraction belonging to or characteristic of an entity
Arabic WordNet ©

attribute predicate محمول
منطقيا - ما يقال على موضوع ومنه القضايا الحملية Propositions attributives
Philosophy Lexicon ©

Philosophy Lexicon
Academy of the Arabic Language-Cairo
Al Amiria Printing House
Academy Portal Browse Online

character | difference | flavor | attribute صفة مميزة

ONTOLOGY

attribute – 3 results (0.04 secs)

▶ Attribute صِفَةٌ | خَصْلَةٌ | مُسْنَدٌ | حالَةٌ | نَعْتٌ | مَزِيَّةٌ | مَيْزَةٌ | شَكْلٌ | هَيْئَةٌ | سِمَةٌ
An abstract that describes qualities of an entity without using units of measure
مجرّد يصف ويُعبر عن قيمة خاصيّة لشيء ما دون استعمال وحدة قياس
example: اللون البرتقاليّ هو صفة لخاصيّة اللون في البرتقال
293263 TypeOf : {Abstract}

▶ Physical Attribute صِفَةٌ مَادِيَّةٌ
Attribute that describes an abstract quality of things
صفة تعبّر عن قيمة خاصية مجرّدة لشيء ما
example: كريم هي صفة مجرّدة لخاصيّة العطاء عن الإنسان
293264 TypeOf : {Attribute}

▶ Abstract Attribute صِفَةٌ مُجَرَّدةٌ | صِفَةٌ إِعْتِبارِيَّةٌ
Attribute that describes an abstract quality of things
صفة تعبّر عن قيمة خاصية مجرّدة لشيء ما
example: كريم هي صفة مجرّدة لخاصيّة العطاء عن الإنسان
293265 TypeOf : {Attribute}

1022977
BIRZEIT UNIVERSITY
Copyright © 2018 Birzeit University

Fig. 3. Screenshot of the search engine's frontend.

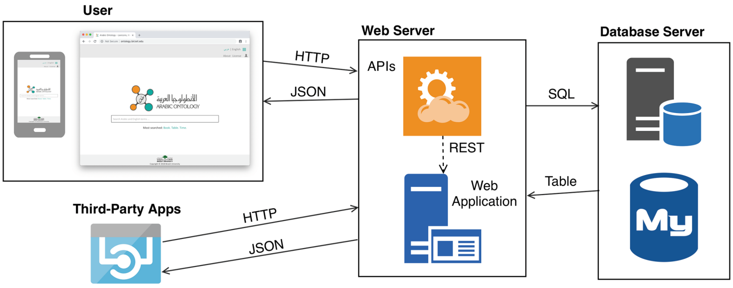


Fig. 4. Search Engine Architecture.

a separate lexical entry or among a synset . In order to keep the URLs Cool, simple and stable, URL parameters (e.g., filters and page number) are passed internally without treating them as part of the URL, e.g. http://ontology.birzeit.edu/term/virus

- *Lexical Concept:* Each lexical concept in all lexicons is given a URL based on its unique *LexicalConceptID*: `http://{domain}/lexicalconcept/{lexicalConceptID}`

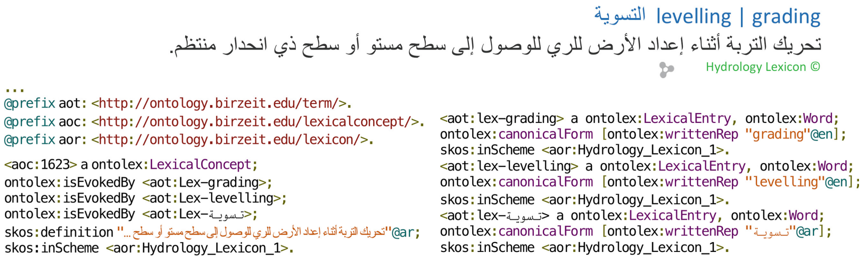
- *Ontology Concept:* Each concept in the Arabic Ontology has a ConceptID and can be accessed using: `http://{domain}/concept/{ConceptID | Term}`. In case of a term, the set of concepts that this term lexicalizes are all retrieved. In case of a ConceptID, the concept and its direct subtypes are retrieved, e.g. http://ontology.birzeit.edu/concept/293198
- *Semantic relations:* Relationships between concepts can be accessed using these schemes: (i) the URL: `http://{domain}/concept/{RelationName}/{ConceptID}` allows retrieval of relationships among ontology concepts. (ii) the URL: `http://{domain}/lexicalconcept/{RelationName}/{lexicalConceptID}` allows retrieval of relations between lexical concepts. For example, http://ontology.birzeit.edu/concept/instances/293121 retrieves the instances of the concept `293121`. The relations that are currently used in our database are: `{subtypes, type, instances, parts, related, similar, equivalent}`.
- *Lemon Representation:* The W3C Lemon representation of each lexical concept in the database is given a URL: `http://{domain}/lemon/lexical concept/{lexicalConceptID}`, e.g. http://ontology.birzeit.edu/lemon/lexicalconcept/1520098340. That is, the RDF representation of any lexical concept can be accessed directly (i.e., not necessarily through the search interface) by adding `/lemon` after the domain in the concept's URL. Additionally, and as illustrated in Fig. 5, an RDF symbol is shown besides each retrieved lexical concept, which links to the Lemon representation of the concept. This is important for referencing and linked data purposes. Nevertheless, this support of Lemon is tentative [10], because of the complexity of treating Arabic lemmas, as noted below. After resolving this, we plan to also provide a Lemon representation of lemmas as well as a SPARQL endpoint for querying data directly.

Remark on Lemma URLs: Each lemma is given a unique LemmaID and a URL: `http://{domain}/lemma/{LemmaID}`, which retrieves the lemma, its morphological features, inflections, and derivations. However, this is partially implemented at this stage (see [10]), as lexical entries in Arabic lexicons are less often lemmas – unlike the case in most English lexicons where a lexical entry is often a lemma (i.e., canonical form). Therefore, each Arabic lexical entry, within the same or across lexicons, needs to be carefully lemmatized first, which is a challenging ongoing task. At this stage, we tentatively consider a lexical entry as a canonical form.

4.2 Presentation of Results

The search engine supports the retrieval of three types of results, each presented in a separate tab, namely Ontology, Dictionaries, and Morphology:

- **Ontology tab**: results in this tab are ontology concepts retrieved only from the Arabic ontology. The tab also allows expanding and exploring the ontology tree.
- **Dictionaries tab**: results in this tab are lexical concepts retrieved from the lexicons. As discussed earlier, a lexical concept can be, for example, a row in a thesaurus (set of synonymous terms), a term(s) and its definition as found in glossaries, or a set of multilingual translations as found in bi/trilingual dictionaries. Figure 5 illustrates a



التسوية levelling | grading

تحريك التربة أثناء إعداد الأرض للري للوصول إلى سطح مستو أو سطح ذي انحدار منتظم.

Hydrology Lexicon ©

```
...
@prefix aot: <http://ontology.birzeit.edu/term/>.
@prefix aoc: <http://ontology.birzeit.edu/lexicalconcept/>.
@prefix aor: <http://ontology.birzeit.edu/lexicon/>.

<aoc:1623> a ontolex:LexicalConcept;
ontolex:isEvokedBy <aot:Lex-grading>;
ontolex:isEvokedBy <aot:Lex-levelling>;
ontolex:isEvokedBy <aot:Lex-تسوية>;
skos:definition "تحريك التربة أثناء إعداد الأرض للري للوصول إلى سطح مستو أو سطح ..."@ar;
skos:inScheme <aor:Hydrology_Lexicon_1>.
```

```
<aot:lex-grading> a ontolex:LexicalEntry, ontolex:Word;
ontolex:canonicalForm [ontolex:writtenRep "grading"@en];
skos:inScheme <aor:Hydrology_Lexicon_1>.
<aot:lex-levelling> a ontolex:LexicalEntry, ontolex:Word;
ontolex:canonicalForm [ontolex:writtenRep "levelling"@en];
skos:inScheme <aor:Hydrology_Lexicon_1>.
<aot:lex-تسوية> a ontolex:LexicalEntry, ontolex:Word;
ontolex:canonicalForm [ontolex:writtenRep "تسوية"@ar];
skos:inScheme <aor:Hydrology_Lexicon_1>.
```

Fig. 5. Example of a lexical concept and its Lemon representation.

lexical concept retrieved from the Hydrology Lexicon. The first line represents the set of synonymous terms in Arabic and English, separated by the symbol "|". The gloss is presented in the second line. The RDF symbol in the third line refers to the Lemon RDF representation of this concept.

- **Morphology tab:** results in this tab are linguistic features, lemma(s), inflections, and derivations of the searched term. This tab is not fully functional yet because our linguistic data is not fully integrated since most dictionaries are not lemmatized.

Additionally, we plan to introduce a forth Dialect tab to allow users to also view the dialectal features [13, 14]. In this way, the four tabs would, to more or less, reflect the different language levels (ontology, meaning, syntax, and dialect).

4.3 Ranking of Search Results

Ranking lexical concepts based on their relevancy was a challenging task. People use lexicographic search for different purposes [4], e.g. searching for translations, definitions, synonyms and/or others. In what follows, we present three ranking strategies: *citation, lexicon's renown*, and a *hybrid* approach which we have adopted.

The **citation strategy ($\mathbf{R_{cit}}$)** ranks each lexical concept based on the frequency of its terms –by counting how many times each of its terms appears as a lexical entry in all lexicons, i.e. the concept's rank is the summation of its terms' frequencies.

$$R = \sum_{n=1}^{|A|} \sum_{m=1}^{k} F_{a_{nm}}$$

$$R_{cit} = \frac{R - R_{min}}{R_{max} - R_{min}}$$

Where, A: is the set of synonyms of a lexical concept, in all languages. k: is the number of lexicons. $F_{a_{nm}}$: is the number of times a_n appears as a lexical entry in lexicon m where $a_n \in A$. R_{cit}: is the citation rank of the concept normalized to be between [0-1].

Our assumption is that the more the concept's terms appear in lexicons, the more this concept is likely to be important. However, its disadvantage is that it decreases the rank of uncommon concepts that are likely to be searched for. Additionally, it scatters the results of the same lexicon across pages, which might confuse users.

The **lexicon renown ranking strategy ($\mathbf{R_{ren}}$)** does not assign a specific rank for each lexical concept. Rather, it assigns each lexicon a rank based on its renown, whether it is general or domain-specific and of high or low quality. Since these criteria are subjective, each lexicon was manually ranked, with respect to other lexicons, by a group of experts. The rank of a lexical concept, then, is given the rank of its lexicon. This allows the renowned results to appear first, but the disadvantage is that linguistic-oriented lexicons are always promoted first, while e.g. the user might be looking for specialized translations. To overcome this, we implemented three types of filters that can be used to show only *translations*, *synonyms*, and/or *definitions*.

The **hybrid ranking strategy ($\mathbf{R_{hyb}}$)** is a combination metric of both strategies above. It ranks the results based on the lexicon renown, and then uses the citation strategy to rank each lexicon's results, which can be obtained by the summation of both ranks:

$$R_{hyb} = R_{ren} + R_{cit}$$

4.4 Usability and Performance Evaluation

We summarize two experiments to evaluate the usability of the search engine, the full details can be found in [4]. First: a subjective **user satisfiability survey** of 12 questions was distributed and answered by 620 users. The answers to these 12 questions are categorized as: efficient (75%), effective (80%), learnable (83%), and good design (73%); and 90% responded that they will use the search engine again. Second: a more objective **controlled usability experiment** was also conducted in a lab, involving 12 users that we arranged into four groups. Each group was given eight tasks to answer. The tasks required users to find synonyms, translations, definitions, and semantic variations between terms. Two groups were asked to use Google Translate, and the other two to use ours. The accuracy of the groups' answers was evaluated by an expert, which was 73% using ours compared to 38% using Google Translate.

Performance: The search engine is currently deployed on a Linux virtual server with average resources (8-core CPU and 16 GB RAM). To estimate its response time (i.e. both backend and frontend processing and retrieval), an experiment was conducted on three user machines. Each was installed in a different location and connected to a different internet service provider. They were programmed to simultaneously send 1 million requests at the rate of 1000 requests/minute, and record the frontend-to-frontend response time for each request. Although the response time is impacted by the network traffic, the experiment showed that it ranged between (0.001 s) and (0.200 s) for all requests.

5 Conclusion and Future Work

We presented a large Arabic multilingual linguistic database, which contains about 150 lexicons of different types, and discussed the different phases carried out to structure, normalize and index this database. We introduced a lexicographic search engine with state-of-the-art design, respecting the W3C's recommendations and best practices.

We plan to continue digitizing more lexicons and adding more functionalities to the search engine, specially the support for French and other languages. Our priority is to lemmatize all lexical entries and then link them across all lexicons. This will enable the interlinking of our lexicographic database with the Linguistic Data Cloud.

Acknowledgments. The authors are thankful to Mohannad Saidi, Mohammad Dwaikat, and other students and former employees who helped us in the technical development and digitization phases. We would like to also thank John P. McCrae for helping us in representing our lexical data in the W3C lemon model. We are also thankful to all lexicon owners, especially the ALECSO team who provided us with many lexicons and supported us in the digitization process.

References

1. Khemakhem, A., Gargouri, B., Hamadou, A.B., Francopoulo, G.: ISO standard modeling of a large Arabic dictionary. Nat. Lang. Eng. **22**(6), 849–879 (2016)
2. Hyland, B., Atemezing, G., Villazón-Terrazas, B.: Best practices for publishing linked data. World Wide Web Consortium (2014)
3. Kamholz, D., Pool, J., Colowick, S.M.: PanLex: building a resource for panlingual lexical translation. In: LREC 2014 (2014)
4. Al-Hafi, D., Amayreh, H., Jarrar, M.: Usability Evaluating of a Lexicographic Search Engine. Technical Report. Birzeit University (2019)
5. Amayreh, H., Dwaikat, M., Jarrar, M.: Lexicons Digitization. Technical Report. Birzeit University (2019)
6. Maks, I., Tiberius, C., Veenendaal, R.V.: Standardising bilingual lexical resources according to the lexicon markup framework. In: LREC 2018 Proceedings (2008)
7. McCrae, J.P., Chiarcos, C., Bond, F., Cimiano, P., et al.: The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. LREC (2016)
8. Helou, M.A., Palmonari, M., Jarrar, M.: Effectiveness of automatic translations for cross-lingual ontology mapping. J. Artif. Intell. Res. **55**(1), 165–208 (2016). AI Access Foundation
9. Jarrar, M.: The arabic ontology - an arabic wordnet with ontologically clean content. Appl. Ontol. J. (2019, Forthcoming). IOS Press
10. Jarrar, M., Amayreh, H., McCrae, J.: Progress on representing Arabic Lexicons in Lemon. In: The 2nd Conference on Language, Data and Knowledge (LDK 2019). Leipzig, Germany (2019)
11. Jarrar, M., Zaraket, F., Asia, R., Amayreh, H.: Diacritic-based matching of Arabic Words. ACM Trans. Asian Low-Resource Langu. Inf. Process. **18**(2), 10 (2018)
12. Jarrar, M., Ceusters, W.: Classifying processes and basic formal ontology. In: The 8th International Conference on Biomedical Ontology (ICBO), Newcastle, UK (2017)
13. Jarrar, M., Habash, N., Alrimawi, F., Akra, D., Zalmout, N.: Curras: an annotated corpus for the Palestinian Arabic Dialect. J. Lang. Resources Eval. **51**(3), 745–775 (2017)

14. Jarrar, M., Habash, N., Akra, D., Zalmout, N.: Building a corpus for Palestinian Arabic: a preliminary study. In: Workshop on Arabic Natural Language Processing (EMNLP 2014). Association for Computational Linguistics (ACL), Qatar, pp. 18–27 (2014)
15. Jarrar, M.: Building a formal Arabic ontology (Invited Paper). In: Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks at ALECSO, Tunis (2011)
16. Jarrar, M., Meersman, R.: Ontology engineering – the DOGMA approach. In: Dillon, T.S., Chang, E., Meersman, R., Sycara, K. (eds.) Advances in Web Semantics I. LNCS, vol. 4891, pp. 7–34. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89784-2_2
17. Jarrar, M., Keet, M., Dongilli, P.: Multilingual verbalization of ORM conceptual models and axiomatized ontologies. Technical report. Vrije Universiteit Brussel (2006)
18. Jarrar, M.: Position paper: towards the notion of gloss, and the adoption of linguistic resources in formal ontology engineering. In: The Web Conference (WWW 2006). ACM (2006)
19. Jarrar, M.: Towards methodological principles for ontology engineering. Ph.D. Thesis. Vrije Universiteit Brussel (2005)
20. Khalfi, M., Nahli, O., Zarghili, A.: Classical dictionary Al-Qamus in lemon. In: 4th IEEE International Colloquium on Information Science and Technology. IEEE (2016)
21. Soudani, N., Bounhas, I., Elayeb, B., Slimani, Y.: An LMF-based normalization approach of Arabic Islamic dictionaries for Arabic word sense disambiguation: application on hadith. J. Islamic Appl. Comput. Sci. **3**(2), 10–18 (2015)
22. Cimiano, P., McCrae, J.P., Buitelaar, P.: Lexicon Model for Ontologies. Final Community Group Report. World Wide Web Consortium (2016)
23. Speer, R., Chin, J., Havasi, C.: ConceptNet 5.5: an open multilingual graph of general knowledge. In: The 31st AAAI Conference on Artificial Intelligence (2016)
24. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. AI 193 (2012)
25. Salmon-Alt, S., Akrout, A., Romary, L.: Proposals for a normalized representation of Standard Arabic full form lexica. In: The International Conference on Machine Intelligence (2005)