Representing Arabic Lexicons in Lemon - a Preliminary Study

Mustafa Jarrar
Birzeit University, Palestine
mjarrar@birzeit.edu

Hamzeh Amayreh
Birzeit University, Palestine
hamayreh@staff.birzeit.edu

John P. McCrae
National University of Ireland Galway, Ireland
john.mccrae@insight-centre.org

Abstract

We present our progress in representing 150 Arabic multilingual lexicons using Lemon, which we have been digitizing from scratch. These lexicons are available through a lexicographic search engine (https://ontology.birzeit.edu) that allows searching for translations, synonyms, and definitions. Representing these lexicons in Lemon will enable them to be used by ontologies and NLP applications, as well as to be interlinked with the Open Linguistic Data Cloud.

2012 ACM Subject Classification Information systems → Resource Description Framework (RDF); Computing methodologies → Language resources

Keywords and phrases Arabic, Lexicographic Database, Lexicographic search engine, Lemon Model

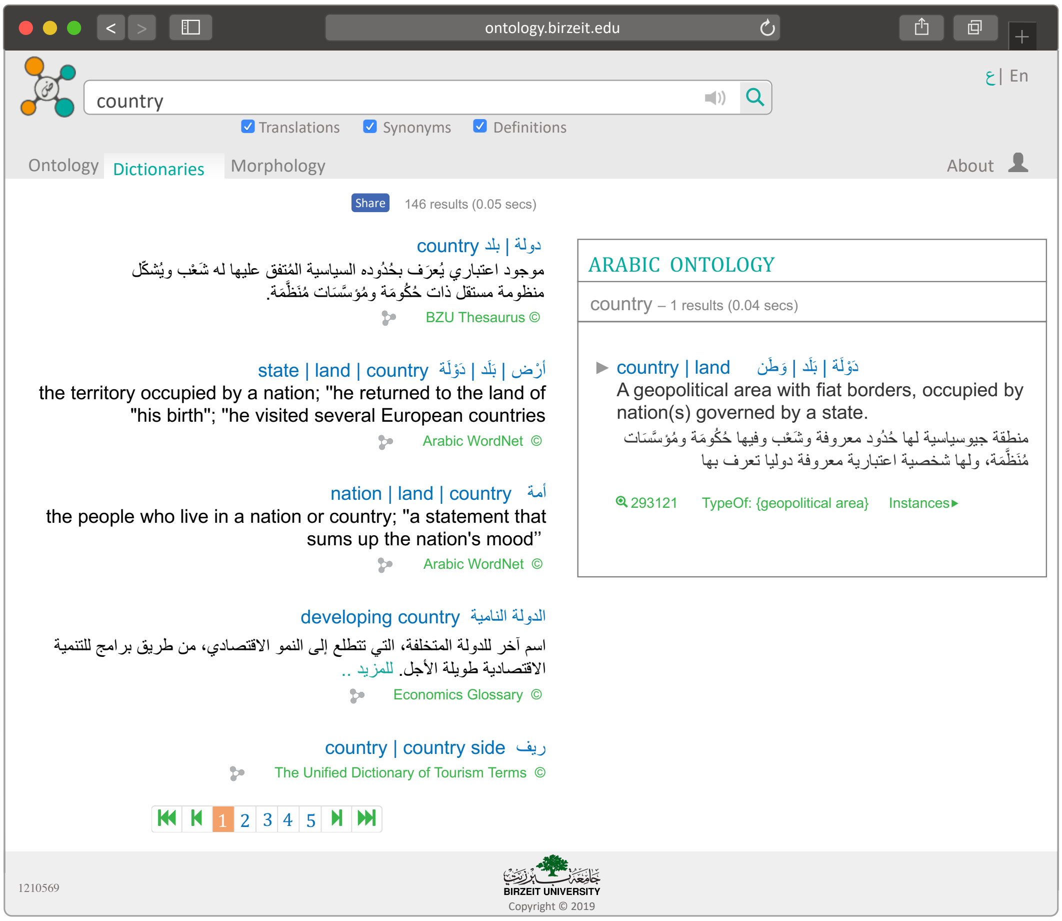
1 Introduction and Related Work

New trends in lexical semantics are demanding lexicons not only to be digitized and well-structured but to also be published and interlinked with other resources. This was realized by the Linguistic Linked Open Data paradigm [13], which is a large collaborative community project to interlink the lexical entries of many different linguistic data sources. The W3C's Lemon RDF model [2], developed by the OntoLex Community Group, aims to enable lexicons to be used by ontologies and NLP applications [4]. Lemon can be used to describe lexical entries and their syntactic and semantic information, encouraging not only the reuse of existing lexicographic data inside modern IT applications, but also the interlinking with other lexicographic resources. Unlike many languages, there is only a limited number of structured Arabic lexicons available in digital format. Earlier attempts to digitize and represent Arabic lexicons using the ISO LMF standard [3] can be found in [14] for Arabic morphological data, [12] for Dutch-Arabic linguistic data, [11] for Al-Madar lexicon, and [15] for classical Hadith lexicons. A recent attempt to digitize Al-Qamus Almuhit lexicon and represent it using the W3C's Lemon can be found in [10]. However, none of these attempts provided access to their lexicons or interlinked it with other resources.

In this paper, we report on our progress on representing 150 Arabic mono/multilingual lexicons using the W3C's Lemon model. These lexicons were digitized over 9 years, during which we had to obtain copyright permissions, digitize most of them by hand, then clean, restructure, normalize and store them in a database – forming the largest Arabic lexicographic database (see [7] and [1]). The database currently contains about 1.1M lexical concepts, 2.4M multilingual lexical entries, 1.5M translation pairs, 0.7M glosses, and 0.5M semantic relations. The database also contains the Arabic Ontology, which is a formal Arabic wordnet that we have built on the basis of a carefully designed ontology [6][5]. It consists currently of about 1.3K concepts, in addition to 11K concepts being validated. The Arabic ontology, which is mapped to WordNet, BFO, and DOLCE, is currently being used

LDK 2019 - Posters Track.
Editors: Thierry Declerck and John P. McCrae

to reference lexical concepts in all lexicons, as will be explained later. A lexicographic search engine [7] was built atop this database (see Figure 1), allowing people to search for translations, synonyms, definitions, morphology, and other information. The results are retrieved from the ontology and the 150 lexicons. As will be explained later, an RDF icon is shown beside each retrieved result (i.e., a lexical concept), which allows accessing the Lemon representation of this concept. The search engine also allows applications to query the database directly, through a set of RESTfull web services, and retrieve the results in JSON format.



country

☑ Translations ☑ Synonyms ☑ Definitions

146 results (0.05 secs)

country بلد | دولة
موجود اعتباري يُعرَف بحُدُوده السياسية المُتفق عليها له شَعْب ويُشكّل
منظومة مستقل ذات حُكُومَة ومُؤسَّسَات مُنَظَّمَة.
BZU Thesaurus ©

state | land | country أَرْض | بَلَد | دَوْلَة
the territory occupied by a nation; "he returned to the land of "his birth"; "he visited several European countries
Arabic WordNet ©

nation | land | country أمة
the people who live in a nation or country; "a statement that sums up the nation's mood"
Arabic WordNet ©

developing country الدولة النامية
اسم آخر للدولة المتخلفة، التي تتطلع إلى النمو الاقتصادي، من طريق برامج للتنمية
الاقتصادية طويلة الأجل. للمزيد ..
Economics Glossary ©

country | country side ريف
The Unified Dictionary of Tourism Terms ©

1 2 3 4 5

ARABIC ONTOLOGY

country – 1 results (0.04 secs)

▶ country | land دَوْلَة | بَلَد | وَطَن
A geopolitical area with fiat borders, occupied by nation(s) governed by a state.
منطقة جيوسياسية لها حُدُود معروفة وشَعْب وفيها حُكُومَة ومُؤسَّسَات
مُنَظَّمَة، ولها شخصية اعتبارية معروفة دوليا تعرف بها
293121 TypeOf: {geopolitical area} Instances▸

Figure 1 Illustration of the lexicographic Search Engine.

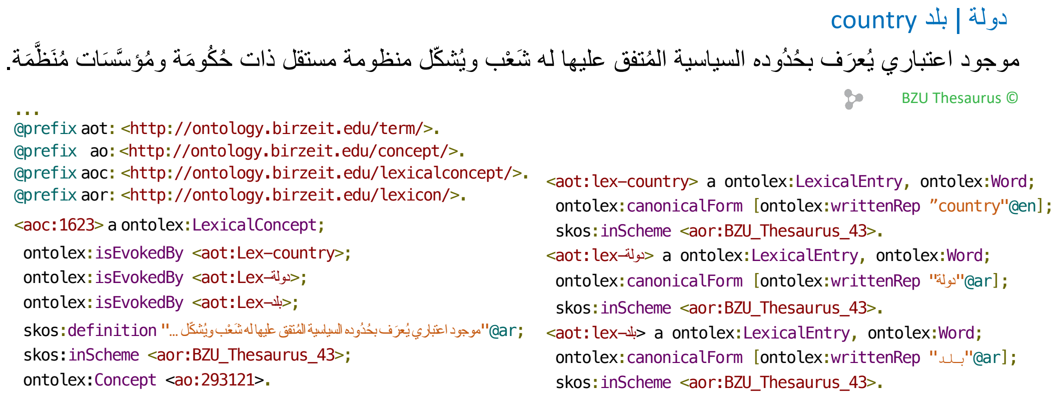
2 Representing Arabic Lexicons in Lemon

Representing Arabic multilingual lexicons using Lemon is non-trivial, because of some specificities of Arabic, and because there are different types of lexicons with different structures. Before discussing these challenges, it is important to understand the types of lexicons according to their internal structure and content type, which we relatively classify as: (i) *Dictionary*: a list of terms, each with some bi/trilingual translations. (ii) *Thesaurus*: sets of synonymous lexical entries, in one or multiple languages, and might contain relations between these sets. (iii) *Glossary*: a set of entries each with a domain-specific short gloss. Advanced glossaries may also provide synonyms, translation(s), and references to other entries, e.g. equivalent, or related. (iv) *Linguistic Lexicon*: a set of headwords, each with its sense(s) and features (e.g., root, POS, and inflections). A headword may have several meanings, which some lexicons designate into separate senses, while in others, senses need to be designated and extracted. (v) *Semantic Variations Lexicon*: a set of pairs of semantically close lexical entries and the differences between them, (e.g. like – love, pain – ache).

In what follows, we present how the content of such types of lexicons is represented in Lemon (illustrated in Figure 2), focusing on Lemon's core semantic features:

- **Lexical entry**: Each translation term in a dictionary, a synonym in a thesaurus, a term in a glossary, or a headword in a linguistic lexicon, is represented as a Lemon's *lexical entry*.
- **Lexical concept**: Each meaning of an entry (a gloss in a glossary, a set of synonyms in a thesaurus, or a translations set in a dictionary) is represented as a Lemon's *lexical concept*. For linguistic lexicons, the different senses of a lexical entry, each is designated and mapped into a separate lexical concept.
- **Ontology concepts**: Each entity in the Arabic Ontology is considered a Lemon's *ontology entity*, and is linked with lexical concepts in other lexicons using the Concept/isConceptOf properties,
- **Relations**: If references to other senses are provided in a lexicon (i.e., semantic relations like *related*, *border/narrower*, etc), we represent them as conceptRel in Lemon.
- **Linguistic features**: Glosses and sense definitions are represented using the skos:definition. Features like POS, root, and inflections are specified using other properties in Lemon.

As illustrated in Figure 1, an RDF icon is displayed beside each of the retrieved results. When this icon is clicked, its lemon representation is generated and shown in a separate page. Figure 2 illustrates the Lemon representation of a lexical concept from the BZU Thesaurus, and its mapping to the concept 293121 in the Arabic Ontology using the Concept property.



دولة | بلد country

موجود اعتباري يُعرَف بحُدُوده السياسية المُتفق عليها له شَعْب ويُشكّل منظومة مستقل ذات حُكُومَة ومُؤسَّسَات مُنَظَّمَة.

BZU Thesaurus ©

```
...
@prefix aot: <http://ontology.birzeit.edu/term/>.
@prefix  ao: <http://ontology.birzeit.edu/concept/>.
@prefix aoc: <http://ontology.birzeit.edu/lexicalconcept/>.
@prefix aor: <http://ontology.birzeit.edu/lexicon/>.
<aoc:1623> a ontolex:LexicalConcept;
 ontolex:isEvokedBy <aot:Lex-country>;
 ontolex:isEvokedBy <aot:Lex-دولة>;
 ontolex:isEvokedBy <aot:Lex-بلد>;
 skos:definition "موجود اعتباري يُعرَف بحُدُوده السياسية المُتفق عليها له شَعْب ويُشكّل ..."@ar;
 skos:inScheme <aor:BZU_Thesaurus_43>;
 ontolex:Concept <ao:293121>.
```

```
<aot:lex-country> a ontolex:LexicalEntry, ontolex:Word;
 ontolex:canonicalForm [ontolex:writtenRep ”country"@en];
 skos:inScheme <aor:BZU_Thesaurus_43>.
<aot:lex-دولة> a ontolex:LexicalEntry, ontolex:Word;
 ontolex:canonicalForm [ontolex:writtenRep "دولة"@ar];
 skos:inScheme <aor:BZU_Thesaurus_43>.
<aot:lex-بلد> a ontolex:LexicalEntry, ontolex:Word;
 ontolex:canonicalForm [ontolex:writtenRep "بـلـد"@ar];
 skos:inScheme <aor:BZU_Thesaurus_43>.
```

Figure 2 Example of a lexical concept and its Lemon representation.

This representation is tentative. Each lexical entry in each lexicon is currently considered a canonical form (i.e., lemma), but in fact it is not always the case. Unlike most English lexicons where a lexical entry is often a lemma, Arabic entries are less often lemmas [9], for two main reasons:

First – many Arabic lexicons do not strictly follow lemmatization conventions. Lemmas are typically used as headwords in lexicons – representing a class of inflectionally related words with the same meanings. In Arabic, the convention for a noun lemma is to be the singular masculine form, and the third person singular perfective form for a verb lemma [8][9]. However, many lexicons are less likely to follow this convention in practice. For example, inflected words like شَارِع (road) and شَوَارِع (roads), يدرك (realizes) and إدراك (realizing), or إدراك (realization) and الإدراك (the realization) might be used as separate headwords within the same or across lexicons. This means that, although such lexical entries are used as separate headwords, they are not necessarily different lemmas. Thus, they should not be considered separate canonical forms when representing them in Lemon. Such cases are more likely to occur in case of dictionaries, glossaries and thesauri. Furthermore, some linguistic lexicons use the same headword for different lemmas. For example, the same word بَيْت could mean (house) with the plural بيوت, and could mean (verse, a piece of poetry) with another plural أبيات. Hence, such ambiguous words should be considered two headwords, or should be given different lemma codes, like (بَيْت 1, بَيْت 2).

Second – lexical entries in Arabic lexicons might be partially or not at all diacritized. Words in Arabic consist of letters and diacritics, thus two words with different diacritics are not necessarily the same word. The problem is that words are typically written without diacritics in practice. This is acceptable by humans who can read and disambiguate words from their contexts, but is more challenging for machines. Headwords in Arabic lexicons might be none or partially diacritized, which makes it difficult to disambiguate them since they have no context, see our experiment in reducing such disambiguation in [9]. For the Lemon representation, considering each headword in a lexicon as a canonical form is not always correct, since headwords might not be fully diacritized; and thus, two none or partially diacritized words might not be the same word within or across lexicons.

To correctly represent Arabic lexical entries in Lemon, each Arabic lexical entry needs to be carefully lemmatized first, which is a challenging task. That is, for each lexical entry, in each of the 150 lexicons, its lemma should be specified. This would enable lexicons then to be interlinked based on their lemmas. Although this is a challenging task as it cannot be fully automated [9], we believe it cannot be avoided specially if lexicons need to be interlinked with external resources as the Linguistic Linked Open Data Cloud.

Furthermore, the Lemon morph module need to be extended to represent some Arabic-specific linguistic and morphological features, such as imperfect and imperative verbs, verbal nouns, intensive participle, place nouns, time nouns, instrumental nouns, and others.

3 Conclusion and Future Work

In this paper, we have presented a tentative representation of 150 Arabic multilingual lexicons using the W3C's Lemon model which can be accessed online. We have discussed the major challenges related to representing lexical entries as canonical forms, especially lemmas and missing diacritics. We plan to conduct a full lemmatization of all lexical entries and disambiguate them in case of missing diacritics. We also plan to extend the Lemon morph module to represent Arabic-specific morphological features.

4 Acknowledgment

This work is part of the Arabic Lemon project funded by the research committee at Birzeit university. Dr. McCrae is supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, co-funded by the European Regional Development Fund, and the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015, ELEXIS - European Lexical Infrastructure and grant agreement No 825182, Prêt-à-LLOD.

References

1 Hamzeh Amayreh, Mohammad Dwaikat, and Mustafa Jarrar. Lexicon Digitization -A Framework for Structuring, Normalizing and Cleaning Lexical Entries, 2018. URL: `https://ontology.birzeit.edu/TR2018.pdf`.

2 Philipp Cimiano, John P. McCrae, and Paul Buitelaar. Lexicon Model for Ontologies: Community Report, 2016. URL: `https://www.w3.org/2016/05/ontolex/`.

3 Gil Francopoulo, Nuria Bel, and et al. Lexical Markup Framework (LMF) for NLP Multilingual Resources. In *Workshop on Multilingual Language Resources and Interoperability*. ACL, 2006.

4 Mustafa Jarrar. Towards the notion of gloss, and the adoption of linguistic resources in formal ontology engineering. In *The 15th international conference on World Wide Web*. ACM Press, 2006.

5 Mustafa Jarrar. Building a Formal Arabic Ontology (Invited Paper). In *Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks*. ALECSO, Arab League, 2011.

6 Mustafa Jarrar. The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. *Applied Ontology Journal*, 2019 [Forthcoming].

7 Mustafa Jarrar and Hamzeh Amayreh. An Arabic-Multilingual Database with a Lexicographic Search Engine. *Proceedings of the 24th International Conference on Applications of Natural Language to Information Systems (NLDB)*, 2019.

8 Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. Curras: An Annotated Corpus for the Palestinian Arabic Dialect. *Journal Language Resources and Evaluation*, 51(3):745–775, 2017.

9 Mustafa Jarrar, Fadi Zaraket, Rami Asia, and Hamzeh Amayreh. Diacritic-Based Matching of Arabic Words. *ACM Asian and Low-Resource Language Information Processing*, 18(2), 2018.

10 M. Khalfi, O. Nahli, and A. Zarghili. Classical Dictionary Al-Qamus in Lemon. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)*, 2016.

11 Aïda Khemakhem, Bilel Gargouri, Abdelmajid Ben Hamadou, and Gil Francopoulo. ISO Standard Modeling of a Large Arabic Dictionary. *Natural Language Engineering*, 22, 2016.

12 Isa Maks, Carole Tiberius, and Remco van Veenendaal. Standardising Bilingual Lexical Resources According to the Lexicon Markup Framework. 01 2008.

13 John McCrae, Christian Chiarcos, Francis Bond, Philipp Cimiano, Thierry Declerck, Gerard de Melo, Jorge Gracia, Sebastian Hellmann, Bettina Klimek, Steven Moran, Petya Osenova, Antonio Pareja-Lora, and Jonathan Pool. The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. 05 2016.

14 Susanne Salmon-Alt, Amine Akrout, and Laurent Romary. Proposals for a Normalized Representation of Standard Arabic Full Form Lexica. In *International Conference on Machine Intelligence*, 2005.

15 Nadia Soudani, Ibrahim Bounhas, Bilel Elayeb, and Yahya Slimani. An LMF-Based Normalization Approach of Arabic Islamic Dictionaries for Arabic Word Sense Disambiguation: Application on Hadith. In *International Conference on Islamic Applications in Computer Science and Technologies*, 2014.