Towards a Sense-based Access to Related Online Lexical Resources

Thierry Declerck, Karlheinz Mörth
DFKI GmbH, Language Technology Lab, Austrian Centre for Digital Humanities
e-mail: declerck@dfki.de, Karlheinz.Moerth@oeaw.ac.at

Abstract

We present an approach aiming at a method to support sense-based cross-dialectal – and cross-lexicon – access to dictionaries of spoken Arabic varieties. The original lexical data consists of three TEI P5 encoded dictionaries describing varieties spoken in Cairo, Damascus and Tunis. This data is included in the Vienna Corpus of Arabic Varieties (VICAV). We briefly present this data, before summarizing the TEI approach for encoding senses in lexical resources. We discuss certain issues related to the TEI representation when it comes to the possibility to provide for a sense-based access to the lexical data. We investigate the use of the recent W3C Ontology-Lexicon Community Group (OntoLex) modeling work and present first results of the mapping of the TEI encoded data onto its *ontolex* model and show how this new representation format can efficiently support cross-dictionary sense-based access.

Keywords: senses; ontolex; TEI; SKOS; Arabic dialects

1 Introduction

We present an approach for ensuring sense-based access to distinct but related online lexical resources. The basis of our work consists of three dictionaries of Arabic varieties developed at different times by different colleagues (Mörth et al. 2015; Procházka & Mörth 2013; Mörth et al. 2014) in the context of the Vienna Corpus of Arabic Varieties (VICAV) project.[1] This data has been encoded in TEI,[2] following the approach described in (Budin et al. 2012). Our current work revises and extends a former approach dealing with a SKOS[3] encoding of the three TEI dictionaries (Declerck et al. 2014), porting it to the use of the W3C model *ontolex,*[4] which has been developed on the basis of Semantic Web technologies. Our focus here is not on the syntactic mapping from selected TEI elements onto *ontolex*, which is straightforward to achieve, but on issues we discovered when taking full advantage of the linking possibilities offered by the *ontolex* model. Our main goal was to support a sense-based access to the lexical resources. Furthermore we describe in this paper the problems we discovered in the TEI encoding of senses of entries, and the solutions we propose to those issues. In the following, we briefly present first the original TEI encoding of the lexicons of Arabic varieties and then discuss the TEI approach to the representation of senses. Finally, we describe the *ontolex* model before presenting our suggestions for supporting a sense-based access to the lexical data.

[1] See https://minerva.arz.oeaw.ac.at/vicav2/. Accessed [30/04/2016].
[2] See (TEI Consortium 2016).
[3] See https://www.w3.org/2004/02/skos/ for more details.
[4] See https://www.w3.org/community/ontolex/wiki/Final_Model_Specification and https://github.com/cimiano/ontolex. Accessed [30/04/2016].

2 The Lexical Data: Three Arabic Varieties Encoded in TEI

The Arabic varieties that have been encoded in TEI P5 are those of the Arab capitals Damascus, Cairo and Tunis.[5] The data on which the dictionaries are based is partly based on interviews and also includes speech transcripts (Mörth et al. 2015). We display below in Figure 1 a full entry from the "Damascus" dictionary all of which correlate with Standard Arabic *bāb* 'door; chapter'. In Figure 2 and 3 we only display the senses associated with the equivalent entries in the "Cairo" and "Tunis" dictionaries; the three entries share the same root consonants "bwb". This use of the root element, reflecting the corresponding roots of Standard Arabic wherever possible, is aiming at facilitating cross-dialectal comparative search. We are currently investigating the use of the sense element for a similar cross-dialectal (and in general cross-lexicon) access to entries.

```
<entry xml:id="baab_001">
  <form type="lemma">
    <orth xml:lang="ar-apc-x-damascus-vicav">bāb</orth>
  </form>
  <gramGrp>
    <gram type="pos">noun</gram>
    <gram type="root"
          xml:lang="ar-apc-x-damascus-vicav">bwb</gram>
  </gramGrp>
  <form type="inflected" ana="#n_pl">
    <orth xml:lang="ar-apc-x-damascus-vicav">bwāb</orth>
  </form>
  <sense>
    <cit type="translation" xml:lang="en">
      <quote>door</quote>
    </cit>
    <cit type="translation" xml:lang="en">
      <quote>gate</quote>
    </cit>
    <cit type="translation" xml:lang="en">
      <quote>city gate</quote>
    </cit>
    <cit type="translation" xml:lang="de">
      <quote>Tür</quote>
    </cit>
    <cit type="translation" xml:lang="de">
      <quote>Tor</quote>
    </cit>
    <cit type="translation" xml:lang="de">
      <quote>Stadttor</quote>
    </cit>
  </sense>
</entry>
```

Figure 1: A "Damascus" entry with root "bwb".

```
<sense>
    <cit type="translation" xml:lang="en"
      <quote>door</quote>
    </cit>
    <cit type="translation" xml:lang="en">
      <quote>gate</quote>
    </cit>
    <cit type="translation" xml:lang="en">
      <quote>gateway</quote>
    </cit>
    <cit type="translation" xml:lang="de">
      <quote>Tür</quote>
    </cit>
    <cit type="translation" xml:lang="de">
      <quote>Tor</quote>
    </cit>
</sense>
```

Figure 2: Sense for a "Cairo" entry with root "bwb"

```
<sense>
    <cit type="translation" xml:lang="en">
      <quote>door</quote>
    </cit>
    <cit type="translation" xml:lang="de">
      <quote>Tür</quote>
    </cit>
    <cit type="translation" xml:lang="fr">
      <quote>porte</quote>
    </cit>
</sense>
```

Figure 3: Sense for a "Tunis" entry with root "bwb".

The focus for the reader should be in those 3 figures on the encoding of the senses for the entries. The sense of each entry is encoded as a set of translations in other languages. No other information is being used for encoding the sense. The reader will observe that there are some differences in the encoded senses, although we would assume that the senses should be identical (or at least similar)

[5] See for more details https://minerva.arz.oeaw.ac.at/vicav2/query/glossary. Accessed [30/04/2016].

across the three entries. In Figure 1 we have a list of six translations equally distributed over 2 languages, English and German. In Figure 2 we also have three English translations but only 2 German translations. Comparing the English translations in Figure 1 and 2, we notice that only 2 of them are identical. The entry in Figure 3 has 3 translations in three different languages, English, French and German. The question is thus on how we can obtain a representation of these different encodings of senses so that we can state that they are identical, or at least having some semantic relation, so that querying by senses all the different dictionaries would lead to the corresponding entries.

3 The TEI Approach to Encoding of Senses in Dictionaries

The TEI approach to encoding of senses is described in chapter 9 "Dictionaries" of the TEI Guidelines (TEI Consortium 2016), which is dedicated to the representation of lexical resources. There, an entry is defined as a component-level element (tagged as <*entry*>) that "contains a single structured entry in any kind of lexical resource, such as a dictionary or lexicon" (TEI Consortium 2016: 276). A sense in TEI (tagged as <*sense*>) is supposed to group "together all information relating to one word sense in a dictionary entry, for example definitions, examples, and translation equivalents." (TEI Consortium 2016: 278). As such a sense is a component of an entry or of elements of an entry, like homonyms. Examples clarifying this view are given in the Guidelines (TEI Consortium 2016: 279). So, for an entry with two senses the following abstract construct (Example 1) is proposed. For the case of an entry with a homograph and two senses, we display here a simplified version of the TEI example (Example 2).

(1)
```
<entry>
  <sense n="1"/>
  <sense n="2"/>
</entry>
```

(2)
```
<entry>
  <hom n="1">
      <sense n="1"/>
      <sense n="2"/>
  </hom>
</entry>
```

A TEI sense can include <*usg*>, <*def*>, <*cit*> elements, whereas the <*cit*> element "contains a quotation from some other document, together with a bibliographic reference to its source. In a dictionary it may contain an example text with at least one occurrence of the word form, used in the sense being described, or a translation of the headword, or an example." (TEI Consortium 2016: 280). In Figures 1-3 in Section 2, the sense information is marked using the <*cit*> element, since the senses are introduced by translations of the headword.

There are no defined restrictions as to how to codify the content of the sense, and all possible string characters seem to be allowed.[6] In our opinion this fact renders the comparison of senses across lexicons difficult, if not impossible. In general we would not like to have to rely on string matching for stating a relation between senses included in different entries in different dictionaries.

[6] Although we regard the above quoted form with the distinct <quote> elements (Figure 1) as current best practice, albeit traditionally lumping together several related senses in one quote (e.g. <cit><quote>door, gate</quote></cit>) is quite common too,

Considering the issue of updates to the senses and their maintenance, it is also not practicable to have the description of a sense distributed over different entries. So for example in Figure 3 a French translation has been added to the set of translations. In case this sense is (nearly) identical to the senses listed in Figure 1 and Figure 2, one would have to add this French translation for those entries as well. It would be preferable to have one sense described in a repository and to include there all the relevant translations. All entries of the different dictionaries would then point to such a repository and so make explicit if different entries are sharing a sense.

Being aware of the recent development of standards for lexical markup in both XML and RDF, we turned our attention to the latter format and described a sense repository in the context of the specification of the *lexicon model for ontologies* (lemon) which results from the work of the W3C Ontology Lexicon Community Group, and more specifically to the core module of this specification, the *ontolex* model.[7] In the next section we present this model, before presenting our current proposal for encoding shared senses for entries of the dictionaries of Arabic variants we briefly described in Section 2.

4 The Ontolex Model

The *ontolex* model has been designed using the Semantic Web formal representation languages OWL, RDF(S) and RDF.[8] It also makes use of the SKOS vocabulary. *ontolex* has been inspired by the ISO Lexical Markup Framework (Francopoulo et al. 2006), which also has an XML serialisation. *ontolex* describes a modular approach to lexicon specification. All elements, sense inclusive, of a lexicon are described independently, while they are connected by typed relation markers. The components of each lexicon entry in the core module are linked by OWL, RDF(s), RDF, SKOS and *ontolex* properties, as this can be seen in Figure 4. The boxes represent classes in the model, and arrows represent specific relations (*properties*) that can be defined between (instances of) classes. The main motivation for the development of *ontolex* was to support the specification of the meaning of lexical entries by pointing to objects described in ontological frameworks, using for this the properties ontolex:denotes or ontolex:reference, offering thus a bridge – or interface – between knowledge of words and knowledge of the world. Lexical senses can play in this bridging a central role.

We wrote scripts for porting (subsets of) the original TEI lexicons, for the time being only for the Cairo and Damascus dictionaries, into *ontolex*. In this encoding we have for the time being 5139 objects as instances of the *ontolex* class "LexicalEntry", 15926 morphological form variants, encoded as instances of the class "Form" and 10189 objects encoded as instances of the LexicalSense class. Examples of senses associated with a lexical entry are given below in Example 7 and Example 8 after the *ontolex* encoding of the entry (Example 3) and related information:

(3) ontolex:apc_eng_baab_001
rdf:type ontolex:LexicalEntry ;
ontolex:otherForm ontolex:form_Inflected_0_apc_eng_baab_001 ;
ontolex:canonicalForm ontolex:form_Lemma_0_apc_eng_baab_001 ;
lexinfo:partOfSpeech lexinfo:noun ;
ontolex:otherForm ontolex:form_Root_0_apc_eng_baab_001 ;

[7] See https://www.w3.org/community/ontolex/wiki/Final_Model_Specification and (McCrae et al. 2012).

[8] See respectively http://www.w3.org/2001/sw/wiki/OWL, http://www.w3.org/TR/rdf-schema/ and http://www.w3.org/RDF/

```
    ontolex:sense ontolex:sense_Sense_0_apc_eng_baab_001 ;
    ontolex:sense ontolex:sense_Sense_1_apc_eng_baab_001 ;
.
```

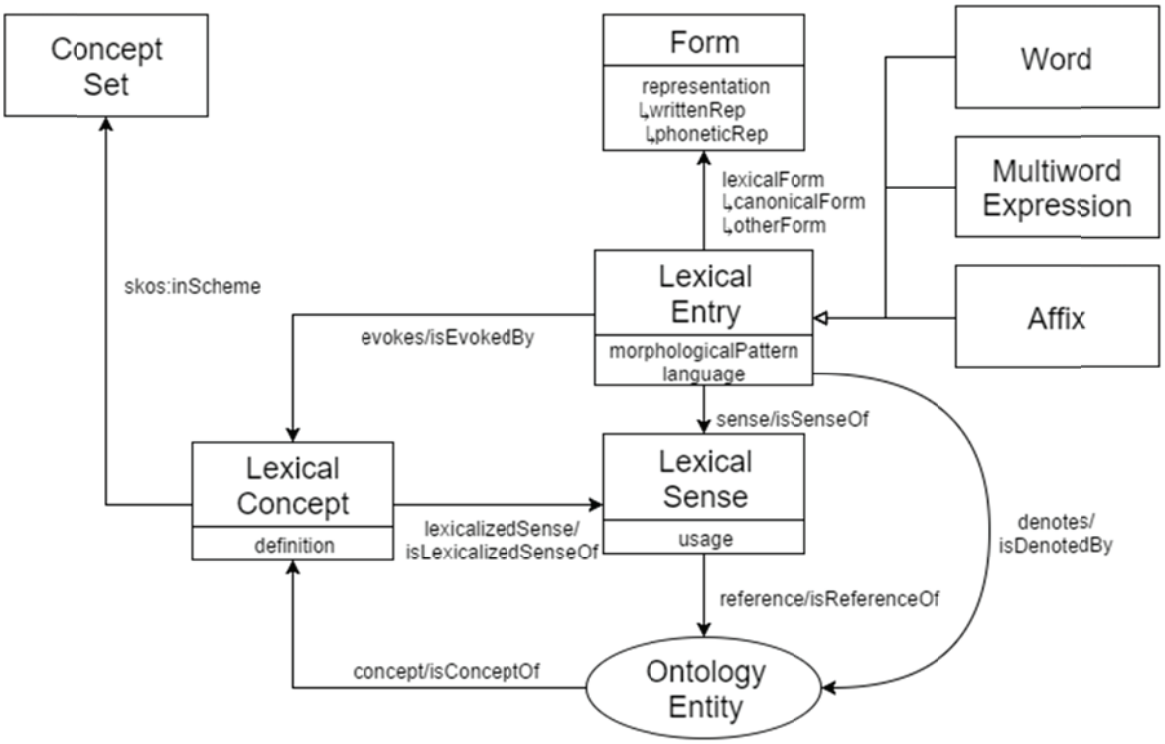

Figure 4: The core model of OntoLex. Figure created by John P. McCrae for the W3C OntoLex Community Group.

```
(4) ontolex:form_Root_0_apc_eng_baab_001
      rdf:type ontolex:Form ;
      ontolex:writtenRep "bwb"^^xsd:string ;
      ontolex:language "ar-apc-x-damascus-vicavTrans"^^xsd:string ;
    .
(5) ontolex:form_Inflected_0_apc_eng_baab_001
      rdf:type ontolex:Form ;
      ontolex:writtenRep " bwāb"^^xsd:string ;
      ontolex:language "ar_apc_x_damascus_vicavTrans"^^xsd:string ;
    .
(6) ontolex:form_Lemma_0_apc_eng_baab_001
      rdf:type ontolex:Form ;
      ontolex:writtenRep "*bāb"^^xsd:string ;
      ontolex:language "ar_apc_x_damascus_vicavTrans"^^xsd:string ;
    .
(7) ontolex:sense_Sense_0_apc_eng_baab_001
      rdf:type ontolex:LexicalSense ;
      rdfs:label "door, gate, city gate"^^xsd:string ;
      ontolex:language "en"^^xsd:string ;
    .
(8) ontolex:sense_Sense_1_apc_eng_baab_001
      rdf:type ontolex:LexicalSense ;
      rdfs:label "Tür, Tor, Stadttor"^^xsd:string ;
      ontolex:language "de"^^xsd:string ;
    .
```

The explanation of this RDF/ontolex code is: The lexical entry “baab” (3, from the “Damascus”

lexicon) has a canonical form (6), some morphological forms, among those the "root" element (4), and 2 senses (7 & 8). All those descriptions are independent "objects" that can be (partially) linked to each other by typed relations (or *properties*). Our mapping from TEI onto *ontolex* is of purely syntactic nature in this case, we have two sense objects corresponding each to a language specific "translation" feature included in the original TEI "sense" element (see Fig. 1).

The lexical entry "bab_001" of the Cairo dictionary has a very similar representation,[9] repeating there the information concerning the root "bwb" and the translations into English and German. This is not only an unnecessary reduplication of information; it is also obvious that this representation lacks the option to access similar entries in the various dialectal dictionaries.

Therefore there is a need to adapt the representation of the results of the purely syntactic mapping.

5 Requirements for a Sense-based Access to the Lexical Data

The requirements to be met by the lexicon developer is therefore to provide for the senses also a unique reference identifier (URI) in the RDF format, so that all lexical entries in the various lexicons can point to a unique object and in this manner support sense-based access to all the lexical data. Expressions in different languages as well as form variants can be linked to this URI.

Encoding the meaning/sense of entries by means of an object uniquely referenced by a URI is not only beneficial for the example displayed in section 4. It also helps to ensure consistency of the encoding of senses, which in the original TEI lexicons was very deficient. This fact is also due to the situation that there is no consistency check provided by this in the used TEI code. And if one takes into consideration that the filling of the "sense" slots in the original TEI lexicons was performed by different persons at different points of time and in the context of different projects, it is not surprising to find many inconsistencies. Just to give a few examples: one sense was encoded with this string "to cover with clay ???", in which case the encoder was marking her/his doubts. For marking options, some encoders use comma, or "/", or just write "or" between the alternatives. This could be easily avoided if the encoders can select from a set of available sense objects, or suggest the creation of new sense objects (to be encoded as instances of the *ontolex* class "LexicalSense"). Via the reference capabilities of the model, one can also point to already existing repositories of senses, being the RDF version of WordNet or other lexical data available in the Linguistic Linked Open Data Cloud,[10] like BabelNet.[11]

Our modifications to the internal organisation of the original TEI encoded data is simple: We generate independent instances of the class "Form" and of the class "LexicalSense" (see Figure 4) for elements of the original TEI lexicons that are shared across various entries. The instances of the class "Form" are representing the "roots" of the entries, and the instances of the class "LexicalSense" are representing the senses. An example for each type of instance is given below in Examples 9 and 10.

(9) ontolex:Form_Root_bwb
rdf:type ontolex:Form ;
rdfs:comment "The root form of various entries in Arabic varieties"^^xsd:string ;
rdfs:label "\"bwb\""@ar ;
ontolex:language "Standard Arabic"^^xsd:string ;

[9] We note also that in both the Damascus and the Cairo dictionary we also have the entry "bawwaaba_001" with root "bwb" and translations "gate" (en) and "Tor" (de). There is also the need to be able to describe sense-relations within one dictionary.
[10] See http://linguistic-lod.org/llod-cloud.
[11] See http://babelnet.org/.

```
        ontolex:representation "bwb"@ar ;
    .
```

Each instance of LexicalEntry having this root information is just pointing to this instance, which is thus introduced only one time in the *ontolex* lexicon. The same remark is valid for the example 10, but there we are dealing with senses.

```
(10)  ontolex:LexicalSense_Door
        rdf:type ontolex:LexicalSense ;
        rdfs:comment "A sense for the original TEI entries containing \"door\" in their sense"^^xsd:string ;
        rdfs:label "Tür"@de ;
        rdfs:label "door"@en ;
        rdfs:label "porte"@fr ;
        ontolex:reference < https://www.wikidata.org/wiki/Q36794 > ;
    .
```

Access to all the instances of LexicalEntry sharing one root or one sense is ensured by simple SPARQL queries, as shown in Figure 5 and Figure 6 below. In Figure 5 we display all the entries that are linking to the root data "bwb", while in Figure 6 we show the entries that are sharing the sense "Door". We can naturally also address combined searches.

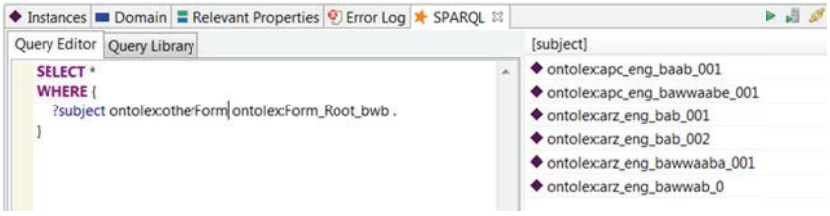

Figure 5: Result of the SPARQL query for entries sharing the root object "bwb".

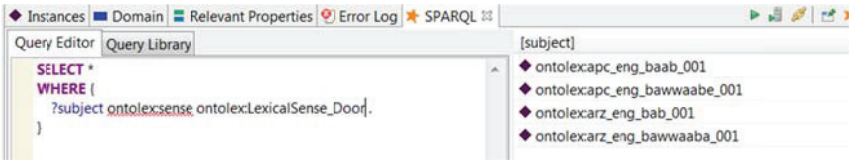

Figure 6: Result of the SPARQL query for entries sharing the sense object "Door".

As the reader can observe in (10), with the use of the *ontolex* model we are in the position to link senses to referential objects representing related encyclopaedic knowledge outside of the lexicon. In this example, we point to the Wikidata entry https://www.wikidata.org/wiki/Q36794, which in turn links (among others) to Wikipedia pages in 96 languages,[12] proposing thus additionally 95 translations of the word "door".

6 Conclusion

In our report, we described some requirements for the encoding of senses in different but related dictionaries in order to support an effective sense-based access to entries of such lexical resources. Those requirements led to a mapping from TEI encoded lexicons into *ontolex* structures. In addition, the mapping was also meant to be used in support of linking of lexical resources to objects available

[12] Like the Arabic Wikipedia page for "door":
https://ar.wikipedia.org/wiki/%D8%A8%D8%A7%D8%A8_%28%D8%AF%D8%A7%D8%AE%D9%84%D9%8A%29

in the (Linguistic) Linked Data Cloud. However, to achieve this end, manual work will not do. Therefore we are currently investigating the application of automatic processes to support lexicographers in this task.

7 References

Budin, G., Majewski, S. & Mörth, K. (2012). Creating Lexical Resources in TEI P5. In *Journal of the Text Encoding Initiative,* 3 (doi:10.4000/jtei.522).

Budin, G. & Mörth, K. (2011). Hooking up to the corpus: the Viennese Lexicographic Editor's corpus interface. In I. Kosem & K. Kosem. (eds.) Electronic lexicography in the 21st century: new applications for new users, *Proceedings of eLex 2011 conference*, Bled, Slovenia: Trojina, Institute for Applied Slovene Studies, pp. 52-59.

Declerck, T., Mörth, K. & Wandl-Vogt, E. (2014). A SKOS-based Schema for TEI encoded Dictionaries at ICLTT, In *Proceedings of the, Ninth International Conference on Language Resources and Evaluation (LREC 2014),* Reykjavik, Iceland.

Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M. & Soria, C. (2006). Lexical Markup Framework (LMF). In *Proceedings of the fifth international conference on Language Resources and Evaluation.*

McCrae, J.-P., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, P., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D. & Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. In *Journal for Language Resources and Evaluation*, 46(4), pp. 701-719.

Moerth, K., Prochazka, S. & Dallaji, I. (2014). Laying the Foundations for a Diachronic Dictionary of Tunis Arabic. A First Glance at an Evolving New Language Resource. In A. Abel, Vettori, C. & Ralli, N., *Proceedings of the XVI EURALEX International Congress: The User in Focus* (pp. 377-387). Bolzano / Bozen, Italy.

Mörth, K., Schopper, D. & Siam, O. (2015). Towards a diatopic dictionary of spoken Arabic varieties: challenges in compiling the VICAV dictionaries, *In Proceedings of the 11th Conference of AIDA - Association Internationale de Dialectologie Arabe*, Bucharest, Romania.

Procházka, S., & Mörth, K. (2013). The Vienna Corpus of Arabic Varieties: building a digital research environment for Arabic dialects. In *Proceedings of the 10th AIDA Conference*, Doha, Qatar.

TEI Consortium 2016. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.0.0. Last updated on 29th March 2016, revision 89ba24e. Accessed at http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf . Accessed [30/04/2016].

Acknowledgements

Work presented in this paper has been supported by the PHEME FP7 project (grant No. 611233), the FREME H2020 project (grant No. 644771) and the Tunico FWF Project (grant No. P 25706-G23).