A SKOS-based Schema for TEI encoded Dictionaries at ICLTT

Thierry Declerck[1], Karlheinz Mörth[2], Eveline Wand-Vogt[2]
[1] DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg, 3, D-66123 Saarbrücken, Germany
[2] Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences
Sonnenfelsgasse 19/8, A-1010 Vienna, Austria
declerck@dfki.de, {Karlheinz.Moerth|Eveline.Wandl-Vogt}@oeaw.ac.at

Abstract

At our institutes we are working with quite some dictionaries and lexical resources in the field of less-resourced language data, like dialects and historical languages. We are aiming at publishing those lexical data in the Linked Open Data framework in order to link them with available data sets for highly-resourced languages and elevating them thus to the same "digital dignity" the mainstream languages have gained. In this paper we concentrate on two TEI encoded variants of the Arabic language and propose a mapping of this TEI encoded data onto SKOS, showing how the lexical entries of the two dialectal dictionaries can be linked to other language resources available in the Linked Open Data cloud.

Keywords: Dialectal dictionaries, TEI, Linked Open Data

1. Introduction

In the context of work recently pursued at ICLTT[1] on porting (German) dialectal dictionaries[2] of the Austrian Academy of Sciences onto the SKOS [3] format (Wandl-Vogt & Declerck, 2013), we started to study the possibility of also mapping TEI[4] encoded dictionaries of Arabic dialects into SKOS, aiming ultimately at a unique SKOS schema that can be used for encoding all electronic dictionaries available at ICLTT. This paper concentrates on actual work consisting in porting to SKOS two dictionaries of Arabic dialects, encoded in TEI and called "ar-apc-x-damascus" and "ar-arz-x-cairo". The building and update of those dictionaries are done in the context of the VICAV project[5] at ICLTT, and the approach implemented for gathering data from the Web and correcting/adjusting these data with the help of NLP resources is described in (Mörth et al., 2013). The final aim of our work is to publish our different dictionary data in the Linked Open Data cloud[6], more specifically in the emerging Linguistic Linked Open framework[7].

2. SKOS

Based on the Resource Description Framework (RDF)[8], SKOS (Simple Knowledge Organization System) [9] "provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary."[10]

Our experiment with SKOS it thus kind of novel, since we apply it to dictionaries, although one can for sure consider dictionaries as being very close to thesauri, and in our approach we encode every entry of the dictionaries as a concept being part of a concept scheme (the dictionary). We chose this representation language, since SKOS concepts can be (1) "semantically related to each other in informal hierarchies and association networks", (2) "the SKOS vocabulary itself can be extended to suit the needs of particular communities of practice" and finally, because it (3) "can also be seen as a bridging technology, providing the missing link between the rigorous logical formalism of ontology languages such as OWL and the chaotic, informal and weakly-structured world of Web-based collaboration tools."[11]

With the use of SKOS (and RDF), we are also in the position to make our dictionary resources compatible with other language resource available in the LOD cloud. Examples of such resources are the actual DBpedia instantiation of Wiktionary[12] or the recent new release of BabelNet[13], both resources encoded using RDF and the *lemon* model[14],which has been developed in the context of the Monnet project[15]. *lemon* is also available as an ontology[16], which we plan to utilize, if appropriate, in a next development step.

3. The Transformation from TEI to SKOS

In this section, we describe briefly the mapping we

[1] ICLTT stands for "Institute for Corpus Linguistics and Text Technology", see http://www.oeaw.ac.at/icltt/

[2] More specifically the "Dictionary of Bavarian dialects of Austria", see http://www.oeaw.ac.at/dinamlex/WBOE.html

[3] See http://www.w3.org/TR/skos-primer/ and (Miles et al., 2005)

[4] See http://www.tei-c.org/index.xml and (Romary, 2009)

[5] VICAV stands for "Vienna Corpus of Arabic Varieties". See http://www.oeaw.ac.at/icltt/node/59

[6] See http://linkeddata.org/

[7] http://linguistics.okfn.org/resources/llod/

[8] http://www.w3.org/RDF/

[9] http://www.w3.org/2004/02/skos/

[10] http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/

[11] Ibid.

[12] See http://dbpedia.org/Wiktionary. There, *lemon* is also used for the description of certain lexical properties.

[13] http://babelnet.org/

[14] *lemon* stands for "Lexicon Model for Ontologies". See http://lemon-model.net/ and McCrae et al. (2012)

[15] See www.monnet-project.eu

[16] See http://www.monnet-project.eu/lemon

propose from the TEI encoding onto a SKOS scheme, which we also populate with the information included in the dictionary. Figure 1 below displays an entry from the "Damascus" dictionary

```
<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>
</entry>
```

Figure 1: An entry from the "damascus" dictionary, in the TEI encoding.

Our first step in the SKOS modelling consisted in creating a ConceptScheme:

```
skos:icltt_dictionaries
    rdf:type skos:ConceptScheme .
```

All further concepts used in our SKOS model are encoded as being part of this ConceptScheme. Dictionaries are introduced as sub-classes of the class "Book". Now we show below how the two dictionaries "Damascus" ("ar-apc") and "Cairo" ("ar-arz") are introduced and put into relation using the corresponding SKOS elements (skos:related).

Both dictionaries are typed as SKOS collections and also as icltt:dictionary. We establish an underspecified relationship between both lexicons, whereas this relation can be specified in the future.

```
icltt:ar-apc
    rdf:type
        skos:Collection ,
    icltt:Dictionary ;
    rdfs:label
        "vicav_Damaskus"@de ,
        "vicav_Damascus"@en ;
    skos:inScheme
        skos:icltt_dictionaries ;
    skos:member icltt:baab_001 ;
    skos:related icltt:ar-arz .
```

```
icltt:ar-arz
    rdf:type
        skos:Collection ,
    icltt:Dictionary ;
    rdfs:label
        "vicav_Kairo"@de ,
        "vicav_Cairo"@en ;
    skos:inScheme
        skos:icltt_dictionaries ;
    skos:member icltt:bab_001 ;
    skos:related icltt:ar-apc .
```

We introduce entries of the lexicons via the skos:member property. The names of the objects reflect the original ID in the TEI encoding (see Figure 1 above for the ar-apc case). Entries are complex objects, as there also in the original TEI format. Entries are complex objects, as there also in the original TEI format.

```
icltt:baab_001
    rdf:type icltt:Entry , skos:Concept ;
    icltt:hasForm
        icltt:baab_001_P ,
        icltt:baab_001_A1 ;
    icltt:hasRoot
        "bwb"@ar-apc-x-damascus-vicav ;
    icltt:hasSense
        icltt:door ,
        icltt:city_gate ,
        icltt:gate ;
    skos:inScheme skos:icltt_dictionaries ;
    skosxl:altLabel icltt:baab_001_A1 ;
    skosxl:prefLabel icltt:baab_001_P .
```

```
icltt:bab_001
    rdf:type skos:Concept , icltt:Entry ;
    icltt:hasForm
        icltt:bab_001_P ,
        icltt:bab_001_A1 ;
    icltt:hasRoot
        "bāb"@ar-arz-x-cairo-vicav ;
    icltt:hasSense
        icltt:city_gate ,
        icltt:gate ,
        icltt:door ;
    skos:inScheme skos:icltt_dictionaries ;
    skosxl:altLabel icltt:bab_001_A1 ;
    skosxl:prefLabel icltt:bab_001_P .
```

In the examples of an entry for each dictionary, shown above, the reader can see that we encoded the TEI element "form" of the original entries as one object, which can have various instantiations. The one ending with letter "P" (standing for skosxl:prefLabel) is representing the original "lemma" type, and the one with the ending "A" (standing for skosxl:latLabel), plus an integer, is representing the original "inflected" type. Since there are entries in the dictionaries, which have more than one inflected form, we had an integer for each of the alternative labels. An important aspect of this representation is the fact that both entries are sharing the same senses (those expressed by lemmas in English, German and/or French). And contrary to the pure TEI encoding, we can here take advantage of the possibility to encode the senses as unique objects:

```
icltt:door
        rdf:type icltt:Sense , skos:Concept ;
        rdfs:label "door"@en , "Tür"@de ;
        skos:inScheme skos:icltt_dictionaries .
```

Adding just the reverse property „isSenseOf" to this "sense" object, we get then all the entries that share this "sense", and we can thus easily semantically link entries of distinct dictionaries. Actually we are abstracting about the string representation of the sense, adopted primarily from the original entry in the TEI encoding, and give as the range of the property "hasSense" the corresponding URL of the sense, if available, in the DBpedia instantiation of Wiktionary, which are in this case:

- http://wiktionary.dbpedia.org/page/door-English-Noun-1en (http://wiktionary.dbpedia.org/page/T%C3%BCr-German)
- http://wiktionary.dbpedia.org/page/gate-English (http://wiktionary.dbpedia.org/page/Tor-German-Noun-1de)
- http://wiktionary.dbpedia.org/page/city_gate-English (http://wiktionary.dbpedia.org/page/Stadttor-German-Noun-1de)

The interesting fact here, is that depending on the level of completeness of the description of the senses in the DBpedia instantiation of Wiktionary, we can have access to a certain number of translations, which can be retrieved automatically and linked to the entries of our SKOS representation of the dialectal varieties of Arabic. In doing so, we can link a large number of entries via shared senses.

For the sake of completeness, we display the "leaves" of the SKOS representation of the entries, limiting ourselves to the skosxl:prefLabel cases (representing the "lemma" type of the original entries). In the corresponding boxes below, the reader can see for each dictionary the written representation of the entry itself. For reason of space, we do not present the skosxl:altLabel instances (corresponding to the original "inflected" form types, but we mention that we are aiming here at using the ISOcat data category registry[17] for pointing to values for POS and morphological features.

```
icltt:baab_001_P
        rdf:type
                icltt:Form ,
                skos:Concept ,
                skosxl:Label ,
                icltt:lemma ;
        skos:inScheme skos:icltt_dictionaries ;
        skos:related icltt:bab_001 ;
        skosxl:literalForm
                "bāb"@ar-apc-x-damascus-vicav .
```

```
icltt:bab_001_P
        rdf:type
                icltt:lemma ,
                icltt:Form ,
                skos:Concept ,
                skosxl:Label ;
        skos:inScheme skos:icltt_dictionaries ;
        skosxl:literalForm
                "bāb"@ar-arz-x-cairo-vicav .
```

4. Conclusion

In this paper we described on-going work on the "skosification" of two TEI encoded dictionaries of dialectal variations of Arabic. We show how this leads to the possibility of linking entries from different dictionaries, using for example the "senses" that are common to entries of the two dictionaries. But we show also how this strategy leads to the possibility of linking the entries of the dictionaries to senses encoded in the DBpedia instantiation of Wiktioanry. In doing so, we get the possibility to link to the corresponding set of multilingual entries in Wiktionary. Once we publish in the LOD the SKOS version of the two dialectal dictionaries, other language resources in this framework can also link to the entries to our dictionaries. Next step in our work will consist in analyzing the opportunity to use the *lemon* model, which is based on the ISO LMF standard[18], for encoding more complex entries, consisting in more than one token. Acknowledgements

Place all acknowledgements (including those concerning research grants and funding) in a separate section at the end of the article.

5. Acknowledgements

The DFKI research work on SKOS described in this paper has been co-financed by the European Commission, in the context of the FP7 ICT project TRENDMINER, under contract number 287863.

[17] http://www.isocat.org/

[18] See http://www.lexicalmarkupframework.org/ and (Francopoulo, 2013)

6. References

Chiarcos, C., Cimiano, P., Declerck, T., McCrae, J.P. (2013). Linguistic Linked Open Data (LLOD) - Introduction and Overview. In: Christian Chiarcos, Philipp Cimiano, Thierry Declerck, John P. McCrae (eds.): *2nd Workshop on Linked Data in Linguistics, Pages i-xi,* Pisa, Italy.

Declerck, T., Lendvai, P., Mörth.K. (2013) Collaborative Tools: From Wiktionary to LMF, for Synchronic and Diachronic Language Data. In Francopoulo, G. (ed) *LMF Lexical Markup Framework.* Wiley 2013.

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

Morth, K., Procházka, S., Siam,O., Declerck T. (2013) Spiralling towards perfection: an incremental approach for mutual lexicon-tagger improvement. In: *Proceedings of eLex 2013,* Tallinn, Estonia.

Moulin, C. (2010) Dialect dictionaries - traditional and modern. In: Auer, P., Schmidt, J.E. (2010) (eds) Language and Space. An International Handbook of Linguistic Variation. Volume 1: Theories and Methods. Berlin / New York. pp: 592-612. (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science / Manuels de linguistique et des sciences de communication 30.1).

Miles, A., Matthews, B., Wilson, M. D., Brickley, D. (2005) SKOS Core: Simple Knowledge Organisation for the Web. In *Proc. International Conference on Dublin Core and Metadata Applications*, Madrid, Spain,

Romary, L. (2009). Questions & Answers for TEI Newcomers. *Jahrbuch für Computerphilologie 10*. Mentis Verlag,

Schreibman, S. (2009) The Text Encoding Initiative: An Interchange Format Once Again. *Jahrbuch für Computerphilologie 10*. Mentis Verlag.

Wandl-Vogt, E. (2005) From paper slips to the electronic archive. Cross-linking potential in 90 years of lexicographic work at the Wörterbuch der bairischen Mundarten in Österreich (WBÖ). In: *Complex 2005. Papers in computational lexicography*. Budapest: 243-254.

Wandl-Vogt, E. and Declerck, T. (2013). Mapping a Traditional Dialectal Dictionary with Linked Open Data. In *Proceedings of eLex 2013,* Tallinn, Estonia.