How to semantically relate dialectal Dictionaries in the Linked Data Framework

Thierry Declerck
University of Saarland
Computer Linguistics Department
Postach 15 11 50
D-66041
declerck@dfki.de

Eveline Wandl-Vogt
Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences.
Sonnenfelsgasse 19/8, A-1010 Vienna
Eveline.Wandl-Vog@oeaw.ac.at

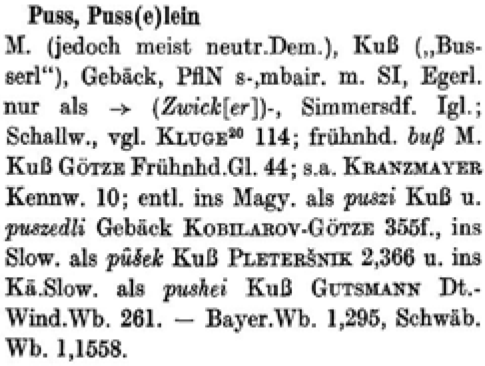
Abstract

We describe on-going work towards publishing language resources included in dialectal dictionaries in the Linked Open Data (LOD) cloud, and so to support wider access to the diverse cultural data associated with such dictionary entries, like the various historical and geographical variations of the use of such words. Beyond this, our approach allows the cross-linking of entries of dialectal dictionaries on the basis of the semantic representation of their senses, and also to link the entries of the dialectal dictionaries to lexical senses available in the LOD framework. This paper focuses on the description of the steps leading to a SKOS-XL and *lemon* encoding of the entries of two Austrian dialectal dictionaries, and how this work supports their cross-linking and linking to other language data in the LOD.

1 Introduction

The starting point for our work is given by two Austrian dialectal dictionaries: The Dictionary of Bavarian dialects of Austria (*Wörterbuch der bairischen Mundarten in Österreich*, WBÖ)[1] and the Dictionary of the Viennese dialect (*Wörterbuch der Wiener Mundart*, WWM)[2]. Both dictionaries have been made available to us in an electronic version: WBÖ in a proprietary XML schema and WWM in Microsoft Word. We used the TEI "OxGarage"[3] service to convert the WWM Word document into a TEI compliant XML representation. Table 1 below shows partially an example of an entry in the printed version of WBÖ.

Table 1: An example for an entry in the WBÖ



Puss, Puss(e)lein
M. (jedoch meist neutr.Dem.), Kuß („Busserl"), Gebäck, PfN s-,mbair. m. SI, Egerl. nur als → (*Zwick*[*er*])-, Simmersdf. Igl.; Schallw., vgl. KLUGE^{20} 114; frühnhd. *buß* M. Kuß GÖTZE Frühnhd.Gl. 44; s.a. KRANZMAYER Kennw. 10; entl. ins Magy. als *puszi* Kuß u. *puszedli* Gebäck KOBILAROV-GÖTZE 355f., ins Slow. als *pûšek* Kuß PLETERŠNIK 2,366 u. ins Kä.Slow. als *pushei* Kuß GUTSMANN Dt.-Wind.Wb. 261. — Bayer.Wb. 1,295, Schwäb. Wb. 1,1558.

In a previous work we ported elements of WBÖ onto SKOS[4] in order to be able to publish entries

[1] http://verlag.oeaw.ac.at/Woerterbuch-der-bairischen-Mundarten-in-Oesterreich-38.-Lieferung-WBOe

[2] See (Hornung & Grüner, 2002).

[3] See http://oxgarage.oucs.ox.ac.uk:8080/ege-webclient/

[4] "SKOS - Simple Knowledge Organization System - 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. As an application of the Resource Description Framework (RDF), SKOS allows concepts to be composed and published on the World Wide Web, linked with data on the Web and integrated into other concept schemes."

Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014, pages 9–12,
Gothenburg, Sweden, April 26 2014. ©2014 Association for Computational Linguistics

of this dictionary in the Linked Data[5] cloud (Wandl-Vogt & Declerck, 2013). We used recently a similar approach for porting the TEI Version of the WWM dictionary into SKOS, leading to few modifications in our previous model.

A motivation for this additional step was to investigate if our SKOS-based model can support the (automatised) cross-linking of the dialectal dictionary data[6]. In this particular case, we can take advantage of a property of dialectal dictionaries concerning the expression of meanings of entries: Although conceived as monolingual reference works, dialectal dictionaries share with bilingual dictionaries the fact that they express the meanings of their entries in a different language. The meta-language for expressing the meanings of entries in both WBÖ and WWM is Standard German, sometimes accompanied by Austrian German. This is exemplified in the WBÖ entry "Puss" in Table 1 above, which is using both the Standard German "Kuß" and the Austrian German "Busserl" for expressing one meaning of the word "Puss" (this meaning being "kiss"). Other meanings are "Gebäck" and "PflN"[7]. Additional lines for the entry "Puss" in WBÖ, not displayed in this submission due to lack of space, are giving more details on those meanings, précising that in the "Gebäck" case we deal with a small sweet pastry ("Kl. süßes Gebäck") and in the "PflN" case with a "bellis perennis" flower.[8]

The related entry in WWM dictionary is "Bussal", which is displayed in Table 2.

Table 2: The related entry in the WWM dictionary

Bussal, Bussi, Bussl, das, 1) Kuss (Syn.: *Schm$tss*); 2) kleines Süßgebäck; Pl. *Bussaln;* viele Komp. wie *Nussbussal* usw. –

We can see that this entry carries two meanings, which are the same as the two first meanings of the WBÖ entry "Puss". Linking entries in distinct dialectal dictionaries can thus be implemented on the basis of meanings that are shared across the dictionaries. But, while for the second meaning the readers familiar with the German language will immediately recognize that both strings "Kl. süßes Gebäck" (WBÖ) and "kleines Süßgebäck" (WWM) have the same meaning, this is not evident for other readers and for computer program that should cross-link the dictionary data from those two sources.

In order to automatically cross-link entries from both dictionaries, we wrote first a program for extracting the strings expressing the meanings for each entry and applied an algorithm for comparing the extracted strings. For this latter task, it is necessary to first linguistically analyse the strings, since pure string matching cannot provide accurate comparisons: lemma reduction and PoS tagging are giving additional indicators for matching strings expressing meanings. To mark linguistically analysed meanings as related, use also semantic representation languages developed in the context of W3C standardization, more specifically SKOS-XL[9] and *lemon*[10]

2 Extraction and Linguistic Analysis of Strings marking Meanings

We wrote for the extraction of strings marking the meanings of entries task specific Perl scripts, adapted to the XML schemas of WBÖ and WWM (in its converted TEI format). Second, we provided an automatic linguistic analysis of those extracted meanings, using lexical and syntactic analysis grammars written with the NooJ finite

(http://www.w3.org/TR/skos-primer/)

[5] For more details see http://linkeddata.org/.

[6] The topic of "cross-linking" is in fact very relevant to lexicographers, as can be seen in (Wandl-Vogt, 2005).

[7] The word "Gebäck" (*pastry*) is Standard German and the string "PflN" is an abbreviation for the German name "Pflanzenname" (*name of a plant*)

[8] More details are given in (Author2 & Author1, 2013). We concentrate in this submission on the sense "small sweet pastry" to exemplify our approach.

[9] http://www.w3.org/TR/skos-reference/skos-xl.html

[10] http://lemon-model.net/ and (McCrae et al., 2012).

state platform[11]. This included tokenization, lemmatisation, Part-of-Speech (POS) tagging and constituency as well as dependency analysis.

The strings marking in both dictionaries the “sweet pastry” meaning are enriched with the following linguistic features:

> WBÖ: (NP süßes (ADJ, lemma = süß, MOD) Gebäck (N, lemma = Gebäck, HEAD))
>
> WWM: (NP (kleines (ADJ, lemma = klein, MOD) Süßgebäck (N, compound: süß (ADJ, lemma = süß, MOD) + Gebäck (N, lemma = Gebäck, HEAD)), HEAD))

In those examples (*sweet pastry* and *small sweet pastry*), we can see the distinct serializations of similar meanings in German. The second example uses a compound noun (“Süßgebäck”), which has the same meaning as the simple nominal phrase in the first example (“süßes Gebäck”). In order to automatically establish this similarity, it is necessary to perform a morphological decomposition of the head noun in the second example. It is also necessary to have the lemma of the adjective in the first example, in order to compare it with the first element of the compound noun in the second example.

The fact, that both linguistically analysed meanings (strings) share the same lemmas for adjectival modifiers and head nouns is the basis for cross-linking the entries. This cross-linking has to be expressed in Semantic Web standards (e.g. compatible to RDF) in order to be published in the Linked Data cloud.

3 Porting the Dictionary Data into the Linked Open Data framework

3.1 Porting the dictionaries into SKOS

Analogue to the described SKOSification of WBÖ (see Wandl-Vogt & Declerck, 2013), the WWM was ported into SKOS. Departing from the former experiment, we decided to not encode anymore the whole dictionary as a SKOS concept scheme. Rather we introduce the listing of entries (each encoded as a skos:Concept) as being member of a skos:Collection. Complementary to this, extracted senses (see former section) are each encoded as skos:Concept being included in a skos:ConceptScheme. This decision is due to the fact that the senses can be organized along the line of (SKOS) semantic relations, whereas the strings marking the entries are in fact just member of a list, which is building the dictionary. The headword (string) of the dictionary entries is encoded as a value of the SKOS-XL prefLabel property. Alternative strings (like “Bussi” in the WWM example in Table 2) are encoded with the SKOS-XL altLabel property. The use of SKOS-XL allows us to “reify” the value of the range of the label properties, and thus to have there not only a literal but further information, like PoS. Since senses are also represented in the dictionaries by strings, we apply the same procedure: a sense has skos-xl labels in which we can encode the lemma of the components of the strings, the corresponding PoS but also related senses, within the local concept scheme or in the LOD, like for example with objects in the DBpedia instantiation of Wiktionary[12].

3.2 Representing the meanings in lemon

The linguistically analysed meanings cannot be (straightforwardly) represented in SKOS, and for this we opted for the *lemon* model, which has been developed specifically for the purpose of representing linguistic information of lexical entries related to knowledge organization systems. The *lemon* encoding of the meanings is incorporated as the value of the SKOS-XL “Label” property. Taking as an example the one meaning of “Puss” in WBÖ that consists of two words (“süßes Gebäck”, *sweet pastry*), we can see that it is for necessary to tokenize the string representing the meaning of the entry “Puss”: the first token can then be lemmatized to “süß” (*sweet*), while for the second token the lemma is identical to the written form used. We represent the

[11] See http://www.nooj4nlp.net/pages/nooj.html

[12] So for example the sense „Kuss“ for both the entries „Puss“ and „Bussal“ is declared as being a skos:exactMatch with the URL: http://wiktionary.dbpedia.org/page/Kuss-German-Noun-1de. From there we can get then all multilingual equivalents listed in this resource.

tokenization information using the *lemon* property “decomposition”.

4 Cross referencing of dictionary entries through similar meanings

The establishment of a relation between “Puss” in WBÖ and “Bussal” in WWM is made possible on the base of the successful mapping of both the adjectival modifier “süß” and the head noun “Gebäck”, which are present in both the definitions in WBÖ and WWM. This similarity is encoded using the “related” property of SKOS. Interesting is also the fact that a user searching the electronic version of the dictionaries could give the High German form “Gebäck” and would get from both dictionaries all the entries which have this word in their definition. The same for the High German adjectival form “süß”.

Instead of the meanings we extracted from the dictionaries, we can use the DBpedia instantiation of Wiktionary as a reference for the senses of the entries of the dictionary, pointing directly to linguistic and knowledge objects that are already in the LOD. Using the “decomposition” and “subSenses” properties of *lemon*, we link to URLs in DBpedia/Wiktionary representing the sense for each token.

5 Conclusion

We described the actual state of RDF/SKOS/lemon modeling of (senses of) entries of dialectal dictionaries, so that those entries can be cross-linked via their similar senses. We have shown that NL processing of the strings for marking the meanings of the entries is necessary in order to make them comparable. We further have shown that our encoding of the entries of the dictionaries is also supporting the linking to already existing lexical senses and other language data in the LOD. The model have been implemented in the TopBraider composer[13] and all the entries of the dictionaries, as instances of the defined classes and properties, are automatically mapped onto the corresponding Turtle syntax[14] and will be made available very soon as deferentiable URLs, making thus less-resourced language data available in the LOD. Future work will consist in applying a similar approach to historical and geographical contexts given in the entries of the dialectal dictionaries.

Acknowledgments

The work by University of Saarland described in this paper is partially supported by the PHEME project, funded by the European Commission under number 611233.

References

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

Hornung, M., Grüner, S. (2002) Wörterbuch der Wiener Mundart; Neubearbeitung. öbvhpt, Wien.

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.

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,

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.

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.

Wörterbuch der bairischen Mundarten in Österreich (WBÖ) (1963-). Wien. Accessed at http://hw.oeaw.ac.at/wboe/31205.xml?frames=yes (25.5.2)

[13] http://www.topquadrant.com/tools/IDE-topbraid-composer-maestro-edition/

[14] http://www.w3.org/TeamSubmission/turtle/