Cross-linking Austrian dialectal Dictionaries through formalized Meanings

Thierry Declerck, Eveline Wandl-Vogt
DFKI GmbH, Language Technology Lab, Germany; Austrian Academy of Sciences, ICLTT, Austria
declerck@dfki.de, eveline.wandl-vogt@oeaw.ac.at

Abstract

This paper deals with the formalization of aspects of definitions used in dialectal dictionaries. We focus on the way "meanings" are encoded in such dictionaries, an essential position for many users and lexicographers, and describe how such an encoding can be re-used for cross-linking entries of on-line (dialectal) dictionaries. In this contribution we describe in some details experiments we made in this respect with two Austrian dialectal dictionaries: The dictionary of Bavarian dialects of Austria ("Wörterbuch der bairischen Mundarten in Österreich", WBÖ) and the dictionary of the Viennese dialect ("Wörterbuch der Wiener Mundart", WWM), which we ported into the SKOS and *lemon* models in order to publish them in the Linguistic Linked Open Data cloud. We show how this approach is not only appropriate for supporting the automation of the cross-linking of dialectal dictionaries, but also for linking entries of the dialectal dictionaries to other types of lexical and encyclopaedic resources in the web

Keywords: Dialectal lexicography; Semantic Web; Linguistic Linked Open Data; Austrian dialects

1 Introduction

In the context of recent work dedicated to porting 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] onto representation formats supporting their publication in the Linked Open Data (LOD) framework[3], and more specifically in the Linguistic Linked Open Data cloud[4], we got our attention directed to the investigation on how this approach could support an automation of the cross-linking of such dialectal language resources. For this, we focused on the way "meanings" are encoded in the selected dictionaries, an essential position for many users and lexicographers. We take advantage here of a property of dialectal dictionaries concerning the expression of meanings of entries: Although conceived as monolingual reference work, 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 and Austrian German, as can be seen for example in the WBÖ entry "Puss": one of its meanings is expressed by the Standard German word "Kuß" (*kiss*) and by the Austrian German word "Busserl" (WBÖ: 1,1515)[5], as can be seen in Figure 1 below. WWM uses also the Standard German word "Kuss" for expressing a meaning of the entry "Bussal", as can be seen in Figure 2. Our assumption is thus that linking entries in distinct dialectal dictionaries can be implemented on the base of meanings that are expressed by similar means across the dictionaries.

In this paper we first briefly describe the two dialectal dictionaries we have been considering for our experiments. We then depict the processes we applied to the entries of the dictionaries for extracting and analyzing the expressions that express their meanings, and their encoding in the representation languages RDF[6], SKOS-XL[7] and *lemon*[8] for supporting their publication in the LOD.

The ultimate goal of our work is not only to be able to cross-link the lexical resources described in this paper, but also to link them in the Linked Data cloud with available data sets for highly-resourced languages and to elevate this way our dialectal and historical lexical resources to the same "digital dignity" as the mainstream languages have already gained. We show also how this encoding allows enriching our lexical data with additional lexical information, mainly senses and multilingual variants.

2 The selected dialectal Dictionaries WBÖ and WWM

We describe in this section briefly the main characteristics of the two dialectal dictionaries we selected for conduction our experiments on cross-linking.

---

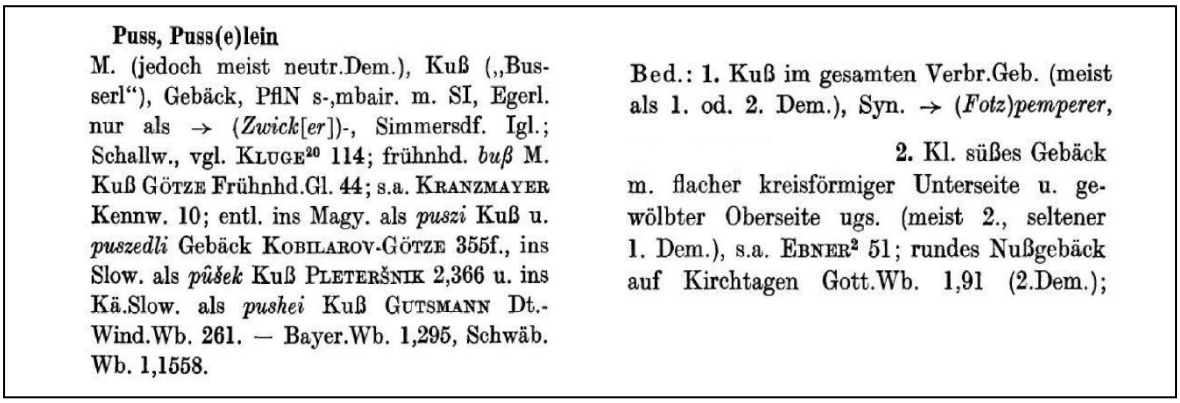
[1] See http://www.oeaw.ac.at/dinamlex/WBOE.html and (Wandl-Vogt 2005; Wandl-Vogt 2008). See also (Declerck & Wandl-Vogt 2013) for a description of the approach adopted for porting WBÖ to the SKOS representation language.
[2] See (Hornung & Grüner 2002).
[3] See http://linkeddata.org/
[4] See http://linguistics.okfn.org/resources/llod/
[5] See ÖWB 141: Bussel das, -s/-[n] (ugs.): Busserl; [ugs.] = colloquial language
[6] See http://www.w3.org/RDF/
[7] See http://www.w3.org/TR/skos-reference/skos-xl.html
[8] See (McCrae & al., 2012).

2.1 The Architecture of the selected dialectal Dictionaries in a Nutshell

The chosen dialectal dictionaries, WBÖ as well as WWM, are scientific dictionaries. Each dictionary offers references and example sentences for illustrating contexts of use for the entries. Whereas the documentation and interpretation in WBÖ is exhaustive, WWM is much shorter and comprehensive. WBÖ is a dialectal dictionary for all Bavarian dialects in the former Austrian Hungarian Monarchy (status: about 1915), whereas WWM is a dictionary for the dialect of the city and county of Vienna.

- grammar: Every entry informs about grammatical properties of the word
- etymology: Every entry contains information about the etymology of the word.
- definition(s): Definitions are a central position in both (onomasiological) dictionaries. Complementing the definitions, WBÖ presents a lot of examples of spoken and written dialect, phrases, songs and poems. Due to the fact, that approximately 10% of the material consists of excerpts of written texts and that the main aim in the beginning was to document the development of a word from its beginning to the actual dialect (see Arbeitsplan 1912) the emphasis on written texts is very high. Furthermore, WBÖ definitions often include a lot of encyclopaedic information about rural traditions and traditional customs.
  WWM presents this type of semantic information in a much more concise way.
  - meanings, as a core part of definitions: Although conceived as monolingual reference work, many 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. So for example the meaning of the WBÖ entry "Puss" is expressed by the Standard German word "Kuß" (*kiss*) and by the Austrian German word "Busserl" (WBÖ 1,1515)[9], as can be seen also in Figure 1. As the reader will see later, we take advantage of this property of dialectal dictionaries concerning the expression of meanings of entries: Linking entries in distinct dialectal dictionaries can be implemented on the base of the Standard German expressions of meanings that are shared across those dictionaries.
- references to dictionaries of adjacent German dialects: WBÖ puts the information presented into a whole dialectal language area in quoting the neighbouring German dialectal dictionaries. WMM does not include this position due to not being embedded into the same / similar methodological background.
- phonetics: Phonetics played an important role in the so called "Junggrammatische Schule"[10], which is the methodological background for the WBÖ. WWM offers headwords that are transliteration based on phonetics.
- compounds: Compounds are treated within the base word entry, e.g. "Nuss)puss" (the famous creamy hazelnut truffle "Nussbusserl"). They might be dealt with in the position 'Komp.' (*Compounds*) within the main entry (as this is the case for the entries *Puss* (WBÖ) or *Bussal* (WWM)).
- cross references to derivations and related words: There is a position, where cross references to derivations and related words are stored, e.g. *Syn.* → *(Fotzen)pemperer* (WBÖ); *(Syn.: Schmåtss)* (WWM). Including articles of derivations or related words, especially those with "less information value" due to a rationalisation concept for WBÖ (see Straffungskonzept 1998: §§ 1.2.1-1.2.3).
- editor: Finally, each WBÖ entry closes with the initial of the author, e.g. *W.B. = Werner Bauer*; in the WWM similar signing is not existing.

In the partial entry for the word "Puss" in WBÖ shown in Figure 1, the reader can see in the right column the details for two selected meanings expressed in Standard German: 1) "Kuss" (*Kiss*) and 2) "Kl. süßes Gebäck" (*small sweet pastry*).



Puss, Puss(e)lein

M. (jedoch meist neutr.Dem.), Kuß („Busserl"), Gebäck, PflN 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.

Bed.: 1. Kuß im gesamten Verbr.Geb. (meist als 1. od. 2. Dem.), Syn. → (*Fotz*)*pemperer*,

2. Kl. süßes Gebäck m. flacher kreisförmiger Unterseite u. gewölbter Oberseite ugs. (meist 2., seltener 1. Dem.), s.a. EBNER[2] 51; rundes Nußgebäck auf Kirchtagen Gott.Wb. 1,91 (2.Dem.);

Figure 1: WBÖ 3,1515f.: Puss, Puss(e)lein.

In the second example, taken from the WWM and displayed in Figure 2, the reader can see that very similar meanings are provided for the entry "Bussal, Bussi, Bussl". While the first meaning is expressed by using exactly the same Standard German word "Kuss" in both cases, the second meaning (*small sweet pastry*) is expressed in each dictionary by using variants: "Kl. süßes Gebäck" vs. "kleines Süßgebäck" .

[9] See ÖWB 141: Bussel das, -s/-[n] (ugs.): Busserl; [ugs.] = colloquial language.
[10] See http://en.wikipedia.org/wiki/Neogrammarian for more details.

Bussal, Bussi, Bussl, das, 1) Kuss (Syn.: *Schmåtss*); 2) kleines Süßgebäck; Pl. *Bussaln;* viele Komp. wie *Nussbussal* usw. –
Etym.: bair.-österr. Schallwort *Puss* Kuss.

Figure 2: WWM 199: Bussal, Bussi, Bussl.

2.2 Access Structure

The main access structure, for both WBÖ and WWM, is the macrostructure, namely the headword.[11]

- WBÖ has chosen due to etymologic-historic considerations a sophisticated, artificial headword, which is difficult to be used as access structure by scientists and in particular causes problems for laypersons. As an example, the German headword "deutsch*"* (*german*) is represented in WBÖ as "teütsch*"*; the Standard German headword "Pflaumenbaum" and the Standard Austrian word "Zwetschkenbaum" (*plum tree*) are represented in the WBÖ as "Zwëtschken)pāum". And a subentry of the main entry "Pāum" (*tree*), the WBÖ headword "Busserl", lacks a standard German representation.
- WWM chooses a transliterated headword, based on phonetics.

So that a cross-referencing and interlinking of dialectal dictionaries, even within the same language area (here: Bavarian variants), does not work without the development of mapping rules. Furthermore, such a mapping would offer just a flat and non-hierarchical interlinking.
This situation motivated the main approach of the work presented here, which consists in investigating if individual word senses, the meanings of entries expressed in Standard and Austrian German words, can serve as an access point for the cross-linking of our two selected dialectal dictionaries, as well as reference point for linking to lexical resources and other knowledge sources in the Web. The following section describes the processes we applied to the entries of the dialectal dictionaries for extracting and analyzing the expressions that express their meanings

3 Extraction and Linguistic Analysis of Expressions introducing the Meanings of Entries

Our first task consisted in detecting and extracting automatically from both dictionaries the strings expressing the core meanings for each entry. Fortunately 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"[12] service to convert the WWM Word document into a TEI compliant XML representation. In both XML representations it was straightforward to describe in Perl scripts the patterns for extracting the meanings of the entries expressed in Standard or Austrian German. But as mentioned at the end of section 2.1, there are discrepancies in the use of Standard or Austrian German word forms across the dictionaries, so that it is often not possible to establish a relation between words expressing meanings across the dialectal dictionaries. Since pure string matching cannot provide accurate comparisons between those expressions, there is the need to apply basic natural language processing to the expressions and to reduce those to their lemmatized form and to mark them up with part-of-speech and morphological information. The comparison of expressions standing for meanings in both dictionaries is then made on the base of such linguistic information associated to the strings. We provided an automatic linguistic analysis of those extracted meanings, using lexical and syntactic analysis grammars written within the SCHUG tools (Declerck 2002). This included tokenization, lemmatisation, Part-of-Speech (POS) tagging and constituency as well as dependency analysis. The strings marking in both dictionaries the "small sweet pastry" meaning are enriched with the following linguistic features:

(1) WBÖ: (NP süßes (ADJ, lemma = süß, MOD) Gebäck (N, lemma = Gebäck, HEAD)) - *sweet pastry*
(2) WWM: (NP (kleines (ADJ, lemma = klein, MOD) Süßgebäck (N, compound: süß (ADJ, lemma = süß, MOD) + Gebäck (N, lemma = Gebäck, HEAD)), HEAD)) - *small sweet pastry*

In the examples (1) and (2), we can see the distinct serializations of similar concepts 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 first perform a morphological decomposition of the head noun in the second example. And we need the lemma of the adjective in the first example, in order to be compared with the first element of the compound noun in the second example.
The fact, that both linguistically analyzed strings expressing the meanings share the same lemmas for adjectival modifiers and head nouns is the base for cross-linking the entries. As we want to formally express this relation, we need to use an appropriate representation language, opting here for Semantic Web standards (e.g. compatible to RDF), also in order to be able to publish our data in the Linked Data cloud.

4 Porting the dictionary data into the Linked Open Data framework

To mark linguistically analyzed meanings as related, it is requested to use semantic web representation languages, like those developed in the context of W3C[13] standardization activities: RDF[14], SKOS, SKOS-XL[15], *lemon*[16]. With this step

[11] Other important positions, navigation and access structures within (dialectal) dictionaries are – or could be – space and time. Geo-referencing with time-stamp stored within a GIS offers possibilities for spatio-temporal visualisation as well as analysis (Wandl-Vogt 2010) and exploratory visually conducted analysis (Theron & Wandl-Vogt 2014).
[12] See http://oxgarage.oucs.ox.ac.uk:8080/ege-webclient/
[13] See http://www.w3.org/.

we want to benefit from the inherent linking (and merging) possibilities offered by Semantic Web languages used in the Linked Data framework, and more specifically we aim at contributing to the emerging Linguistic Linked Open Data cloud[17], integrating dialectal language data into this framework.

4.1 Porting the dictionaries into SKOS

Based on the Resource Description Framework (RDF), 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.[18]

Our experiment with SKOS is 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 elements of entries (basically the meanings) of the dictionaries as concepts being part of a conceptual scheme. We chose this representation language, since:

- SKOS concepts can be "semantically related to each other in informal hierarchies and association networks"[19]
- "the SKOS vocabulary itself can be extended to suit the needs of particular communities of practice"[20]
- SKOS "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."[21]

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[22] or version 2.0 of BabelNet[23].
We decided in the most recent version of our model to encode the strings standing for introducing each entry of a dictionary as a skos:Concept being a member of a skos:Collection, while each associated sense is encoded as skos:Concept that is part of a concept:Scheme. In the first case we deal with a flat list of elements, while in the second case we can model (hierarchical) relations between the meanings (also called "senses"). In the following pages of this section, we present some examples of our model applied to WBÖ, using the so-called turtle serialization.

(3) icltt:Dictionary
rdf:type owl:Class ;
rdfs:comment "Modeling the ICLTT dictionaries"@en ;
rdfs:label "Wörterbuch"@de , "Dictionary"@en ;
rdfs:subClassOf owl:Thing .

We first introduce (ex:3) a "Dictionary" class, of which the WBÖ dictionary is an instance of (ex:4).

(4) icltt:wboe
rdf:type icltt:Dictionary , skos:Collection ;
rdfs:comment "OEAW Dictionary for Bavarian"@en ;
rdfs:label "Wörterbuch der bairischen Mundarten in Österreich"@de , "Bavarian dialects of Austria"@en ;
icltt:hasLanguage icltt:bar ;
skos:member icltt:concept_puss .

As the dialectal dictionaries are encoded as skos:Collection, entries of the dictionaries are modelled as being a skos:member of such collections. The WBÖ entry "Puss" (encoded in ex:4 as the icltt:concept_puss) is therefore listed as a member (for reason of space we display here only one member of the collection). The icltt:concept_puss is introduced in our model as an instance of the class icltt:Entry (ex:5):

(5) icltt:concept_puss
rdf:type icltt:Entry ;
rdfs:label "puss"^^xsd:string .
(6) icltt:Entry
rdf:type owl:Class ;
rdfs:label "Entry"^^xsd:string ;
rdfs:subClassOf skos:Concept .

In doing this we have introduced entries of the WBÖ as a concept being a member of a collection. We still need to introduce in our model the concrete information about the entries, and we describe this step in the next section.

---

[14] See http://de.wikipedia.org/wiki/Resource_Description_Framework.
[15] See http://www.w3.org/2004/02/skos/ and http://www.w3.org/TR/skos-reference/skos-xl.html respectively.
[16] See (McCrae & al., 2012).
[17] See http://linguistics.okfn.org/resources/llod/
[18] http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/
[19] Ibid.
[20] Ibid.
[21] Ibid.
[22] See http://dbpedia.org/Wiktionary. There, *lemon* is also used for the description of certain lexical properties.
[23] http://babelnet.org/

4.2 Representing the Headwords and the Meanings in SKOS-XL and *lemon*

Contrary to most knowledge objects described in the LOD, we do not considers strings (encoding in WBÖ lemma and word forms as part of the language) as being just literals, but as being also knowledge objects. We considered therfore the use of SKOS-XL and of the *lemon* model[24] for representing the string used for headwords and senses. SKOS-XL has proven to be adequate for encoding strings as complex knowledge objects, but not for representing the linguistically analyzed expressions used for marking the meanings. For this we opted for the *lemon* model, which is compatible to SKOS.

(7) icltt:concept_puss
rdf:type icltt:Entry ;
rdfs:label "puss"^^xsd:string ;
skosxl:prefLabel icltt:entry_puss .

In the displayed code in ex:7, the reader can see now the complete representation of the "icltt:concept_puss" object (extending ex:5): the concrete headword in the dictionary is pointed to by the means of the skox-xl property "prefLabel", which contrary to both the rdfs:label and skos:label properties is not having a literal as the possible value of the range of the property, but which is having an object ("icltt:entry_puss", as shown in ex:7) as a value in its range:

(8) icltt:entry_puss
rdf:type icltt:Lemma ;
icltt:hasPos icltt:noun ;
lemon:sense icltt:gebäck , icltt:kuss , icltt:süßes_gebäck ;
skosxl:literalForm "Puss"@bar .

As the reader can observe in ex:8, we can encode the fact that "puss" is a lemma and that it is a noun. More importantly, we can include the senses associated to the WBÖ entry. In fact we are adding a new sense, "icltt:gebäck" (*pastry*). This is a direct consequence of the linguistic analysis of the expression "süßes Gebäck" we described in section 3 (ex:1 and ex:2): since the word "Gebäck" is the head noun in this nominal phrase, we can assume that this head noun is also a meaning (or sense) to be associated to the entry. In doing so we introduce a hierarchical organization of the meanings associated to an entry: "süßes Gebäck" is a specialization of "Gebäck" (using the "skos:broader" relation, as shown in ex:16) . As mentioned above, this is the reason why we use the skos:ConceptScheme construct in order to encode the senses associated to the entries (ex:9-ex:11):

(9) icltt:Senses_ICLTT
rdf:type skos:ConceptScheme ;
rdfs:comment "Senses that are used in ICLTT dictionnaries"@en ;
rdfs:label "Senses"@en , "Bedeuttungen"@de .
(10) icltt:Sense
rdf:type owl:Class ;
rdfs:label "Sense"@en ;
rdfs:subClassOf skos:Concept ;
owl:equivalentClass lemon:LexicalSense .
(11) icltt:kuss
rdf:type icltt:Sense ;
rdfs:label "kiss"@en , "Kuss"@de ;
skos:inScheme icltt:Senses_ICLTT ;
skosxl:prefLabel icltt:sense_kuss .

Lemmas of the expressions used in the original dictionaries for marking meanings are indirectly linked to the class Sense, and are directly instances of the class icltt:Lemma (ex:12-ex:14):

(12) icltt:sense_gebäck
rdf:type icltt:Lemma ;
rdfs:label "Gebäck"@de ;
icltt:hasPos icltt:noun ;
skosxl:literalForm "Gebäck"@de .
(13) icltt:sense_kuss
rdf:type icltt:Lemma ;
rdfs:label "Kuss"@de ;
icltt:hasPos icltt:noun ;
skosxl:literalForm "Kuss"@de .
(14) icltt:sense_süß
rdf:type icltt:Lemma ;
rdfs:label "süß"@de ;
icltt:hasPos icltt:adj ;
skosxl:literalForm "süß"@de .

[24] *lemon* is also available as an ontology: http://lemon-model.net/

We introduce in ex:14 the class “CompoundSense” that allows us to mark the fact that the sense(s) can be resulting from a compound term or a phrase used to express the meaning of an entry.

(15) icltt:CompoundSense
rdf:type owl:Class ;
rdfs:label "Composition of Sense"@en ;
rdfs:subClassOf icltt:Sense .

An instance of such a class is displayed in ex:16, in which the reader can see how we model for the time being the hierarchical relation between the sense “gebäck” and “süßes Gebäck” (using the “skos:broader” relation). We can also encode the fact that the sense of the entry “süß_gebäck” is composed of two senses, but the model needs to be further developed, since it is clear that the sense “süß” cannot be considered only as a sub-sense of “süß_gebäck”, but more as a “modifying” sense. We are currently working on representing with the help of *lemon* such cases of linguistic dependencies.

(16) icltt:süß_gebäck
rdf:type icltt:CompoundSense , lemon:LexicalSense ;
rdfs:label "sweet pastry"@en , "süßes Gebäck"@de ;
lemon:subsense icltt:süß , icltt:gebäck ;
skos:broader icltt:gebäck ;
skos:inScheme icltt:Senses_ICLTT .

In ex:16 we can see the advantage of using a representation model that can encode linguistic properties. In this case, it is for example 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 tokenization information using the *lemon* property “decomposition”, as can be seen in ex:17:

(17) lemon:decomposition
rdfs:domain lemon:LexicalSense ;
rdfs:range icltt:Sense .

For the time being we introduce in our model an explicit listing of components as subclasses of icltt:CompoundSense (see ex:18 and ex:19). The way this encoding is used is shown in ex:21.

(18) icltt:Component1
rdf:type owl:Class ;
rdfs:label ""^^xsd:string ;
rdfs:subClassOf icltt:CompoundSense .
(19) icltt:Component2
rdf:type owl:Class ;
rdfs:label ""^^xsd:string ;
rdfs:subClassOf icltt:CompoundSense .

4.3 Linking to Resources available in the LOD

As the reader can see in the examples (20) and (21) further below, we decided to use the DBpedia instantiation of Wiktionary as a reference for the senses (meanings) of the entries of the dictionary, pointing thus to linguistic and knowledge objects that are already in the LOD. To be more precise, the link to DBpedia/Wiktionary is applied for each token of the expressed meanings. In the case of “süßes Gebäck”, we can thus point to two URLs in DBpedia/Wiktionary, each representing the adequate senses for the actual token. In the name of the URLs used for pointing to DBpedia/Wiktionary we have implicitly also the information about the language and the PoS of the entry. But one could point to the RDF version of ISO data categories[25] for making this information explicit in our model.
Additionally the linking to the appropriate senses in DBpedia/Wiktionary allows accessing all the corresponding multilingual lemmas associated in this resource with a sense. Looking for example at http://wiktionary.dbpedia.org/page/sweet-English-Adjective-1en (corresponing to the URL for the German word, we us in ex:20), we get more than 70 expressions in more than 60 languages.
And the URL http://wiktionary.dbpedia.org/page/pastry-English-Noun-1en refers to ca. 30 expressions in about 25 languages. Having one unique URL for the “sweet” sense of “süß” allows to link all the corresponding entries to a unique reference point, and so to improve comparability of dictionary resources, also at the semantic level.

(20) icltt:süß
rdf:type lemon:LexicalSense , icltt:Component1 ;
rdfs:label "sweet"@en , "süß"@de ;
skos:exactMatch <http://wiktionary.dbpedia.org/resource/süß-German-Adjective-1de> ;
skos:inScheme icltt:Senses_ICLTT .

[25] See http://www.isocat.org/

(21) icltt:gebäck
rdf:type lemon:LexicalSense , icltt:Component2 ;
rdfs:label "Gebäck"@de , "pastry"@en ;
skos:exactMatch <http://wiktionary.dbpedia.org/resource/Gebäck-German-Noun-1de> ;
skos:inScheme icltt:Senses_ICLTT ;
skos:narrower icltt:süß_gebäck ;
skosxl:prefLabel icltt:sense_gebäck .

In the two examples just above, the reader can see how we can link the senses of the entries to existing sources in the LOD. We use for this the skos:exactMatch property (although we could also use *lemon* properties for this). But our model also makes clear that the DBpedia/Wiktionary URL we use for each token is valid only in the context of the compound term we are dealing with. The word "süß" has in DBpedia/Wiktionary more senses, but in the context of "süßes Gebäck" only the one sense that refers to "sweet" is adequate. Using the lemon model allows us thus to disambiguate senses associated to the components of complex terms used in the dictionaries for expressing a meaning.

5 Cross-referencing of Dictionary Entries through shared Meanings

The establishment of a relation between the entry "Puss" in WBÖ and and the entry "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, but used in the context of textual variants, as can be seen in the examples (1) and (2). 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, also if the word is used in a compound form. The same for the High German adjectival form "süß", also irrespectively if this form is inflected or part of a compound word. Our work is thus also addressing in the longer term the semantic access to dialectal dictionaries.

6 Conclusion

We presented an approach consisting in extracting meanings associated to entries in two dialectal dictionaries. Comparison of the expressions used to mark those meaning can be done only after applying basic natural language processing to those expressions. Expressions that are judged as being similar are cross-linked. Furthermore we encode those meanings in Semantic Web representation languages and can so link to lexical and knowledge resources available in the Linked Data Framework. This step in supporting potentially the semantically base cross-linking of our lexical entries to other dialectal dictionaries published in the web.

Current work is dedicated in improving the model for an adequate representation of more complex linguistic phenomena, and also in investigating how our approach could be applied for linking our dictionaries not only to DBpedia/Wiktionary but also to other lexical resources. We think here in particular to portals that already offer a network of dictionaries, like the Trier Wöterbuchnetz[26], which contains a lot of dialectical dictionaries, also offering cross-links between entries. A next step will consist in published the content of our dictionaries in the Web, so that references from other sources in the LOD to our dictionaries can be be implemented.

7 References

Arbeitsplan (1912). Accessed at http://www.oeaw.ac.at/icltt/dinamlex-archiv/Arbeitsplan.pdf [10/04/2014]

Declerck, T. (202). A set of tools for integrating linguistic and non-linguistic information. In: *Proceedings of SAAKM (ECAI Workshop).*

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.

Francopoulo, G. (2013) LMF -- Lexical Markup Framework. Wiley.

Gennari, J., Fergerson, R., Grosso, W. E., Crubezy, M., Eriksson, H. , Noy, N. F. , Tu, S. W., Musen, M. A. (2002). The Evolution of Protégé: An Environment for Knowledge-Based Systems Development

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. (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science / Manuels de linguistique et des sciences de communication 30.1).

Österreichisches Wörterbuch, 42. Auflage. 2012.

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.

[26] http://woerterbuchnetz.de/

Straffungskonzept für das Wörterbuch der bairischen Mundarten in Österreich (WBÖ) (1998). Accessed at http://www.oeaw.ac.at/icltt/dinamlex-archiv/Straffungskonzept_1998.pdf [10.04.2014].

Theron, R., Wandl-Vogt, E. (2014). The fun of exploration: how to access a non-standard language corpus visually. In: *VisLR: Visualization as added value in the development, use and evaluation of LRs. Workshop-Proceedings of LREC2014*.

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. (2008). ..wie man ein Jahrhundertprojekt zeitgemäß halt: Datenbankgestützte Dialektlexikografie am Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX). In: Ernst, P. (ed) 2008, *Bausteine zur Wissenschaftsgeschichte von Dialektologie / germanistischer Sprachwissenschaft im 19. und 20. Jahrhundert. Beiträge zum 2. Kongress der internationalen Gesellschaft für Dialektologie des Deutschen*, Wien: 93-112.

Wandl-Vogt, E. (2010). Point and find: the intuitive user experience in accessing spatially structured dialect dictionaries. *Slavia Centralis* 3 (2010): 35-53.

Wandl-Vogt, E., Declerck, T**.** (2013). Mapping a Traditional Dialectal Dictionary with Linked Open Data**.** In. Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) 2013. *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Wiktionary RDF extraction. Accessed at: http://dbpedia.org/Wiktionary [10/04/2014]

Wörterbuch der bairischen Mundarten in Österreich (WBÖ) (1963-). Verlag der Österreichischen Akademie der Wissenschaften. Wien. Accessed at: http://hw.oeaw.ac.at/cl?frames=yes [10/04/2014].

Acknowledgements

The DFKI part of this work is supported by the LIDER Project. LIDER: "Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe" is a FP7 project with reference number 610782 in the topic ICT-2013.4.1: Content analytics and language technologies.