Towards a Pan European Lexicography by Means of Linked (Open) Data

Thierry Declerck[1], Eveline Wandl-Vogt[2], Karlheinz Mörth[2]

[1] DFKI GmbH, Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany

[2] ACDH-ÖAW, Sonnenfelsgasse 19, 1010 Vienna, Austria

E-mail: declerck@dfki.de, Eveline.Wandl-Vogt@oeaw.ac.at, Karlheinz.Moerth@oeaw.ac.at

Abstract

In the context of the expanding Linked (Open) Data framework (LOD), work has started to encode linguistic resources in the same format as performed for the data sets present in the LOD, and which represent mainly domain specific knowledge. This approach has been extensively discussed in the W3C Ontology-Lexica Community Group, resulting in the "OntoLex" model, and is also being supported by the European LIDER project, leading for example to extensions of the recently created Linguistic Linked Open Data (LLOD) cloud, and by the European FREME project, applying LLOD principles to various industrial use cases. This development is highly relevant to the goals of the European Network of e-Lexicography (ENeL) COST action, and in this respect we performed a number of experiments to encode lexicographic data of various ENeL partners in a LLOD compliant format. We report in this paper on the first steps taken in the cooperation between ENeL and the other aforementioned projects, providing some detail regarding the encoding model we use: OntoLex.

Keywords: e-Lexicography; Linked Open Data; Multilingualism

1. Introduction

In the context of the European Network of e-Lexicography (ENeL) COST action[1] a question we ask is whether a pan European lexicology and lexicography is conceivable. Concerning the potential European lexicology, this question leads us to searching for commonalities in the structure and the concepts used in the various languages of Europe. Therefore, we need to establish a certain level of interoperability in the description of those languages. Are we able for example to detect and markup shared etymologies between European languages, optimally by automatically consulting machine-readable versions of the dictionaries encoding the properties of the languages? Concerning the potential European lexicography, we aim for example to generate multilingual dictionaries on the basis of the shared concepts or meanings that can be detected between digital versions of monolingual dictionaries. For this we need to have access to a standardized representation of the concepts and meanings used in the different dictionaries for describing their entries. By standardized representation we mean the possibility to anchor the various but similar descriptions of meanings for a headword in different dictionaries into a shared and dereferentiable source on the web.

[1] See http://www.elexicography.eu/

Firstly, on this basis, one can attempt to respond to some research questions such as: How many common roots (etymology) are there across European languages, or are there common neologisms[2]? Are there pan European words, or pan European concepts? How to best utilize pan European multilingual corpora[3]? Or how to cross-link, and (partially) merge, the authoritative dictionaries that have been developed over the years by many participants of the ENeL COST action?

The recent development of the Linked (Open) Data (LOD) framework[4] and more specifically of the Linguistic Linked Open Data (LLOD) cloud[5] seem to offer an ideal environment for solving some of the interoperability issues we mentioned above, while also providing a good platform for linking the content of the authoritative dictionaries to other types of data available on the (semantic) web. We present in the next sections the basic ideas of the LLOD framework and the representation model used for publishing and linking language data in this cloud: OntoLex[6].

2. Linguistic Linked Open Data

For this paper we adopt the definition of Linked Data given by Wikipedia: "In computing, linked data (often capitalized as Linked Data) describes a method of publishing structured data so that it can be interlinked and become more useful through semantic queries. It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. This enables data from different sources to be connected and queried"[7]. Data sets that have been published in the Linked Data format can be visualized by the so-called Linked Open Data Cloud diagram[8] or also by other means like the Linked Open Data Graph[9] .

In the context of this further expanding Linked Data framework, work has started to encode linguistic resources in the same format as already existing linked data sets, which primarily consisted of "classical" knowledge objects and entities. In those data sets, language data is mainly used as human readable information encoded for example in the RDF(s) annotation properties "label", "comment" and the like.

[2] One can consider expressions such as "Grexit“ or "Brexit", which seem to be used across Europe.

[3] Here, we consider, for example, the Europarl Corpus (http://www.statmt.org/europarl/)

[4] See http://linkeddata.org/ for more details

[5] See http://linguistics.okfn.org/tag/llod/ for more details.

[6] https://www.w3.org/community/ontolex/

[7] http://en.wikipedia.org/wiki/Linked_data. A more technical definition is given at http://www.w3.org/standards/semanticweb/data

[8] http://lod-cloud.net/

[9] http://inkdroid.org/lod-graph/

Recently, some researchers[10] in the field of Human Language Technology (HLT) and Semantic Web technologies started to work on models and their implementation that would elevate the language data used in existing LOD data sets to the same type of representation as is the case for the encyclopaedic knowledge they were "commenting" and "labelling". Cooperation on those topics has been established between, among others, the Working Group on Open Data in Linguistics[11] and with the European FP7 Support Action "LIDER"[12]. These joint efforts have led to the establishment of a linked data cloud of linguistic resources, which is called Linguistic Linked Open Data (LLOD)[13] and whose data sets are not only linked to other language data sets, but also to the encyclopedic data sets in the LOD. The Linguistic Linked Open Data cloud is also visualized by an online diagram[14], which itself is derived from information contained in the LingHub repository[15] developed in the context of the LIDER project. More recently, cooperation has been established with the H2020 project "FREME" on the automatic enrichment of digital content[16]. In fact, FREME is providing for industrial use cases that are using the LLOD framework. We investigate, in the context of ENeL, if such approaches to LLOD can be applied to authoritative lexicons for (partial) publishing and linking those within this cloud.

The model "OntoLex" is at the core of the publication of language data and linguistic information in the LLOD. This model results from the W3C Ontology-Lexicon community group[17]. Since this model was originally based on LMF[18], which is itself the ISO standard for Natural Language Processing (NLP) lexicons and Machine Readable Dictionaries (MRD), it is an appealing model for lexicographers who are seeking to publish their data in the LOD. In the next section, we briefly present the current state of OntoLex.

3. OntoLex

The OntoLex model has been designed using the Semantic Web formal representation languages OWL, RDFS and RDF[19]. It also makes use of the SKOS and SKOS-XL

[10] See for example Chiarcos et al. (2013a) and Chiarcos et al. (2013b)

[11] See http://linguistics.okfn.org/ for more details.

[12] See http://www.lider-project.eu/ for more details.

[13] See http://linguistics.okfn.org/tag/llod/ for more details.

[14] http://linguistic-lod.org/llod-cloud

[15] See http://linghub.lider-project.eu/. LingHub is an open and domain adapted (semantic) repository for language resources. All metadata are available in standardized Semantic Web representation languages.

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

[17] See also https://github.com/cimiano/ontolex, complementary to https://www.w3.org/community/ontolex/

[18] See (Francopoulo et al., 2006) and http://www.lexicalmarkupframework.org/

[19] See http://www.w3.org/TR/owl-semantics/, http://www.w3.org/TR/rdf-schema/ and http://www.w3.org/RDF/ respectively.

vocabularies[20]. OntoLex is based on the ISO Lexical Markup Framework (LMF) and is an extension of the *lemon* model, which is described in (McCrae et al., 2012). OntoLex describes a modular approach to lexicon specification, thus allowing the e-lexicographer to depart from the "book" view that the headword is the (unique) entry point to information encoded in a dictionary. Senses, usages, concepts, etc. can be independently described, accessed and are all linked to what was considered the headword, and which is now encoded as a virtual entry in a RDF model.

With OntoLex, we can advocate for the fact that all elements of a dictionary entry can be described independently from each other and connected by explicit (typed) relation markers. Now, the components of a dictionary entry can be distributed in a network and linked together by RDF encoded relations/properties. An important aspect of this model is also the relation called "reference". This represents a property that supports the linking of senses of lexicon entries to knowledge objects available in the LOD cloud. This reflects also our view that the meaning of a lexicon (or dictionary) entry is no longer necessarily encoded in the lexicon (or dictionary) but can be referred to in appropriated resources on the (semantic) web.

In practicality, this means that a dictionary author does not need to describe all components or elements of an entry in detail, but that she/he can also draw on existing elements (e.g. the etymology of a word), and can simply refer to it. We are convinced that these properties of the model can facilitate and support the cooperation between scientific lexicographers, and that this can result in virtual and collaborative research environments in the lexicographical field.

Figure 1 below displays the core model of OntoLex[21]. Boxes represent classes of the model. Arrows with filled heads represent object properties, while arrows with empty heads represent the Sub-Class relations. In arrows labeled 'X/Y', X is the name of the object property and Y the name of the inverse property.

[20] SKOS stands for Simple Knowledge Organisation System, see also http://www.w3.org/2004/02/skos/

[21] The figure and the explanations are taken from the wiki page of OntoLex: http://www.w3.org/community/ontolex/wiki/Final_Model_Specification.

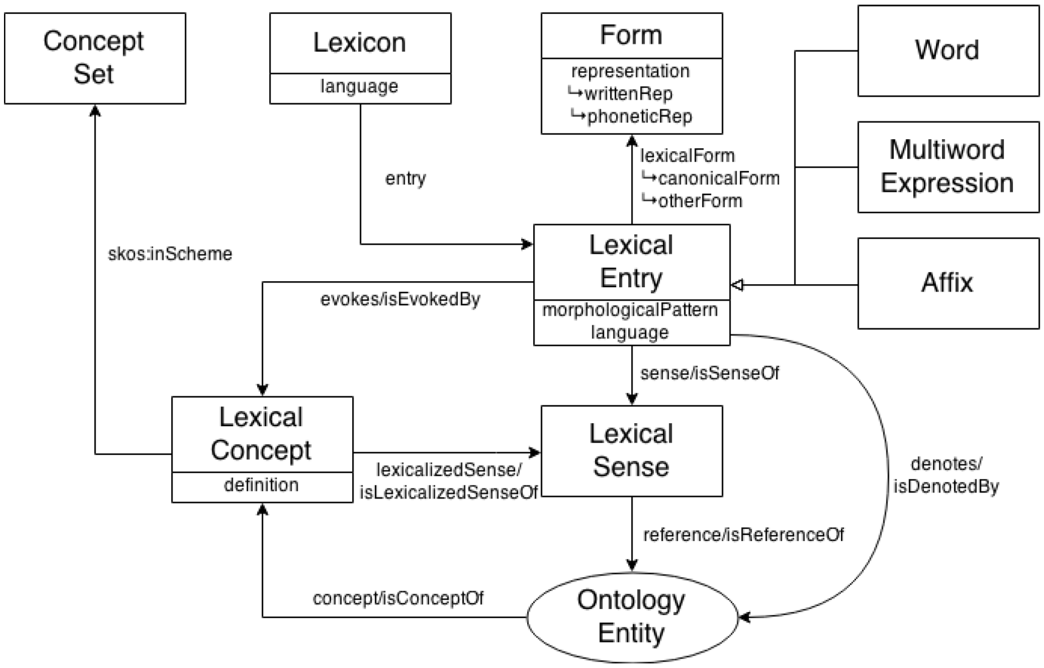

Figure 1: The core model of Ontolex.

Figure created by John P. McCrae for the W3C Ontolex Community Group.

We applied this model on a small list of different types of lexical resources made available by participants of the ENeL network, and we describe this encoding process in the next section.

4. First manual Experiments

In order to test our intuition about the use of OntoLex for the publication of existing authoritative lexicographic resources in the LOD, we provided, as a proof of concept, a manual encoding of some example data provided by ENeL participants in the OntoLex format. The example data we used were taken from:

- 2 Austrian dialect dictionaries (Tustep/XML and Word)
- 1 sample of a Slovak dictionary (XML, + PDF/Word)
- 1 Slovene XML dictionary (XML, based on the LMF standard)
- 2 TEI encoded Arabic dialects (in TEI)
- 1 Sample from a Bask–German dictionary (XML)
- 1 Sample from a French lexicon (extracted from Wiktionary)
- 1 Limburg lexicon (Excel)

- 1 Sample from the KDictionary multilingual source (XML file)
- Sample from the Digital Scottisch Lexicon (Old Scottisch, html + 1 example in TEI)
- 1 Lexicon extracted from a corpus of “Baroque German”

Every dictionary has been encoded in the OntoLex format as an instance of the ontolex:lexicon class, using the ontolex:entry object property to indicate inclusion of an entry.[22] The class ontolex:lexicon thus serves here basically as a container for lexical entries. Below we display the example for the “Wörterbuch der bairischen Mundarten in Österreich” (WBÖ)[23], on which we will focus for the details of the manual encoding in OntoLex[24].

```
  ontolex:WBÖ
   rdf:type ontolex:Lexicon ;
   rdfs:comment "Dictionary of Bavarian Dialects in Austria"@en ;
   ontolex:entry ontolex:lex_trupp ;
   ontolex:entry ontolex:lex_trüllen ;
   ontolex:entry ontolex:lex_trüsche ;
   ontolex:language "bar"^^xsd:string ;
 .
```

In the code displayed above, the reader can see that the lexicon class is acting as a container, in which original entries (here of the WBÖ) are included via the OntoLex property ontolex:entry. The example can be read in natural language as “WBÖ is an instance of the class “Lexicon”, which lists dictionaries and lexicons”. WBÖ deals with the Bavarian Language (“bar”). WBÖ has three entries, “trupp”, “trüllen”, “trüsche”. It is important to note that this instance of a ontolex:lexicon class is indexed by an URI. In our case it is a local one (no longer accessible on the web): http://www.w3.org/ns/lemon/ontolex#wbö. And this is valid for all instances we will see examples of below: they all have an URI, so that their content can be accessed by any sparql queries[25].

In the example above we list only a few examples of entries, as the described experiment was initially performed manually, as a proof of concept.

The entries that are marked in the example of the WBÖ lexicon above in the range of the ontolex:entry object property are themselves instances of the ontolex:LexicalEntry class. The example for the lexical entry “trupp” is displayed below. The lexical entry

[22] All the examples discussed in this section refer to Figure 1.

[23] http://www.oeaw.ac.at/icltt/dinamlex-archiv/WBOE.html

[24] We display all the examples of our OntoLex encoding using the so-called Turtle syntax. Turtle stands for “Terse RDF Triple Language” and is an easily readable serialization of RDF statements. See http://www.w3.org/TR/turtle/ for more details.

[25] SPARQL is a query language defined for RDF triples. See for more details http://www.w3.org/TR/rdf-sparql-query/

`ontolex:lex_trupp` also has some features associated with it, all marked by the use of either datatype or object properties[26]. In the example below, `ontolex:sense` is an example of an object property, while, in the example above, `ontolex:language` is an example of a datatype property.

```
ontolex:lex_trupp
  rdf:type ontolex:LexicalEntry ;
  ontolex:denotes <http://live.dbpedia.org/page/Herd> ;
  ontolex:denotes <http://live.dbpedia.org/page/Social_group> ;
  rdfs:comment "An entry of WBÖ: Trupp"@en ;
  ontolex:canonicalForm ontolex:form_trupp ;
  ontolex:hasEtymology ontolex:ety_trupp ;
  ontolex:sense ontolex:trupp_sense1 ;
  ontolex:sense ontolex:trupp_sense2 ;
  ontolex:sense ontolex:trupp_sense3 ;
.
```

In the example above, we can see that a "canonical from" is defined for the entry. This is due to the fact that OntoLex is supporting the description of variants (regional, typographical, morphological etc.) that are shared by the same entry[27]. In the "lex_trupp" example we can also see how OntoLex deals with semantic ambiguities. There are in this example two usages of the `ontolex:denotes` property. Consulting Figure 1 above, the reader can see that the "denotes" property links directly to an object outside of the "lexical domain". In our case to DBpedia entries, but it could be any domain specific resource. Since we introduced this property twice, we have a clear indication with which we can apply a reference ambiguity. The entry "lex_trupp" also includes three uses of the `ontolex:sense` object property. This property is pointing at objects that are defined as a lexical semantics module within our lexicon space. An example of such a "sense", as an instance of the class "`ontolex:LexicalSense`" is given below.

```
ontolex:trupp_sense1
  rdf:type ontolex:LexicalSense ;
  rdfs:comment "One lexical sense for entry Trupp"@en ;
  ontolex:hasRecord ontolex:rec_trupp1 ;
  ontolex:isSenseOf ontolex:lex_trupp ;
  ontolex:reference <http://live.dbpedia.org/page/Social_group> ;
.
```

As we can see, this object also indicates a DBpedia entry, via the `ontolex:reference` property. The difference between the "denotes" and the "reference" properties is that, in the one case, the domain of the property is an instance of LexicalEntry and, in the second case, it is an instance of the LexicalSense class. In the second case, we can

[26] The distinction between object and datatype properties refers to the fact that a property related to an object can relate either to another object in the ontology (an instance of a class) or to some literal data. See http://www.w3.org/TR/owl-ref/ for more details.

[27] The details of the types of variants currently covered by OntoLex are listed at: http://www.w3.org/community/ontolex/wiki/Specification_of_Requirements/Properties-and-Relations-of-Entries

establish lexical semantic relations between the instances of the class, and this motivates the introduction of this additional referential mechanism.

For both cases, the fact that we can link an entry or, better, a sense to an external resource, like DBpedia, gives access to related multilingual information that is encoded in such a resource. In the case of accessing "`http://live.dbpedia.org/page/Social_group`", we can retrieve related information in many languages (and the potentially related entry in the corresponding language):

- http://fr.dbpedia.org/resource/Groupe_social
- http://de.dbpedia.org/resource/Soziale_Gruppe
- http://cs.dbpedia.org/resource/Sociální_skupina
- http://el.dbpedia.org/resource/Κοινωνική_ομάδα
- http://es.dbpedia.org/resource/Grupo_social
- http://eu.dbpedia.org/resource/Gizarte-talde
- http://id.dbpedia.org/resource/Kelompok_sosial
- http://it.dbpedia.org/resource/Gruppo_sociale
- http://ja.dbpedia.org/resource/社会集団
- http://ko.dbpedia.org/resource/사회_집단
- http://pl.dbpedia.org/resource/Grupa_społeczna

And we also obtain information regarding related Wikipedia categories, like:

- category:Sociology_index
- category:Social_groups
- category:Social_psychology
- category:Sociological_terminology

Looking at the page http://live.dbpedia.org/page/Social_group, the reader can see that there are many other types of information that can be accessed and linked to.

In the first example of the "lex_trupp" entry above, the reader can additionally see that we introduce a property "hasEtymology", which is pointing to an instance of the class "ety(mology)". With this step we further demonstrate how the organization of the digital dictionary can be modularized. All the etymology information contained in the original WBÖ is now contained in a well-defined class of ontology and the instances of this class can be enriched with information from other sources than the WBÖ. The current description of the etymological information included in this WBÖ entry is:

```
ontolex:ety_trupp
    rdf:type ontolex:Etymology_French ;
    rdfs:comment "Instance of a French etymology for the WBÖ entry
\"lex_trupp\" ;
    ontolex:hasCentury 17 ;
    ontolex:hasEtymologyForm "Troupe"@fr ;
    ontolex:isEtymologyOf ontolex:lex_trupp ;
    ontolex:language "French"@en
.
```

This description of the etymology data is very similar to that of the original WBÖ entry "Trupp", which included the etymology in book form. We can create a specific lexicon for all etymological information contained in the WBÖ, and link the entries of this generated etymological lexicon to other etymological resources, and in fact merge all the compatible information. In this way, we are kind of outsourcing some of the information that is not inherently related to the Bavarian dialect to other sources of information that can be more complete and more accurate, since they were put together by real experts in the field of etymology. In doing so, we have a way to compare many lexicographic sources on their shared etymology data, and hence to establish a more complete list of roots that are shared across dictionaries in the LOD format.

A similar remark can be made on the senses (or meanings) of the original entry "Trupp". In the instance `ontolex:trupp_sense1` displayed above, the reader can see that we link this particular sense via the "reference" property to an entry in DBpedia: http://live.dbpedia.org/page/Social_group. From there we can access all dictionaries and other sources that point to this URI, and thus establish a relation with those multilingual resources, accessed from now on by senses or meanings that are represented in DBpedia or in RDF versions of WordNet, and the like.

5. Lessons learned

This section is regarding some lessons learned during our manual OntoLex encoding of (aspects of) various lexicographic resources.

2.1 Representation versus Linking of lexicographical data

It very quickly became apparent that there is no need to provide for an OntoLex based representation of the complete information contained in an original dictionary. As in the case of WBÖ, we can be confronted with quite complex information structures, with different levels of embedding. And since such a dictionary has been developed over a number of years, with many different teams involved, internal consistency of the information and the way it has been encoded is not always given. And in general: the aim is not to propose yet another type of representation but to be able to link (and potentially merge) lexical information. We argue that only this type of information that can be linked should be converted in the OntoLex format and so be published in the Linked (Open) Data framework.

As we know, Tim Berners-Lee outlined four principles of linked data, which are listed on his famous page: http://www.w3.org/DesignIssues/LinkedData.html:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. So that they can discover more things.

We implemented this strategy, but for now limited it to a partial set of the information included in some of the dictionaries we have been working on and in particular the few examples from WBÖ. This limitation is for practical reasons: we so far encoded in OntoLex only the entries, the associated senses and the listed etymology information. This information, available in LOD compliant codes can be linked to related data sets in the Linked Data cloud. If now a user (a human or a machine) wants to access the full amount of information encoded in the WBÖ, we can for example add the full URL of this information under the rdfs:see. Also property to any entry of WBÖ (or other dictionaries) we have been (partially) encoding in OntoLex. Therefore, any data set linking to one of our WBÖ entries encoded in OntoLex will also link to a dereferentiable resource. This will display the original WBÖ entry, as it is encoded in the database version of this dictionary. For example, information about locations that are relevant for an entry can be accessed at http://wboe.oeaw.ac.at/dboe/indices/ort/A/1, etc.

2.2 Manual transformation versus automated transformation

While in this paper we have mainly described a manual work for the OntoLex comprising the encoding of a few (complex) examples from different dictionaries, we also gained some insights into which aspects can be easily automated. If the dictionaries possess clear and consistent structures, so that entries, variants and senses can be easily detected and automatically extracted by means of the applications of patterns expressed as regular expressions in a programming language, automatic OntoLex encoding is possible. It is additionally desirable for the data we obtain to be in a structured format, for example Excel, XML and the like. As an example, we automatically mapped a concept-based lexicon for Limburg dialects, dealing with the anatomy of the human body, from its original Excel format into OntoLex. For this, only some lines of codes were necessary. The original data had 75,355 Excel rows. The lexicon lists in the first column (in a repetitive way) the anatomic concepts (mentioned using standard Dutch language), while in the second and third columns we have the lemma of the dialectal forms and lexical variations of those. The original lexicon is very large, since the concepts of interests are repeated in the first column of the Excel file for every possible variation in the dialect forms, but also for the naming of the different regions in which a variation for the basic concept was found.

After transformation in OntoLex, we have a sense lexicon of only 264 instances. Those correspond in fact to the concepts used in the original lexicon in Excel, and for which 75,355 Excel rows were required. Here, we thus observe the compression power of such a representation in OntoLex (and in RDF in general). In this OntoLex representation, a sense (bovendeel van de rug; *upper part of the back*) has the following form:

```
ontolex:concept_limburg_100
        a   ontolex:LexicalConcept , skos:Concept , ontolex:SenseLexicon ;
        rdfs:comment   "Concept taken from a specific source for the Limburg Language, being
a questionnaire or a dictionary, etc."@en ;
        rdfs:label     "bovendeel van de rug"@nl ;
        ontolex:hasSource    ontolex:source_limburg_4 , ontolex:source_limburg_1 ;
        ontolex:isDenotedBy  ontolex:lex_limburg_239 , ontolex:lex_limburg_1833 ,
ontolex:lex_limburg_1846 , ontolex:lex_limburg_1847 , ontolex:lex_limburg_1826 ,
ontolex:lex_limburg_1834 , ontolex:lex_limburg_1853 , ontolex:lex_limburg_1828 ,
ontolex:lex_limburg_1816 , ontolex:lex_limburg_1829 , ontolex:lex_limburg_1841 ,
ontolex:lex_limburg_1845 , ontolex:lex_limburg_1840 , ontolex:lex_limburg_1831 ,
ontolex:lex_limburg_1844 , ontolex:lex_limburg_1832 , ontolex:lex_limburg_1824 ,
ontolex:lex_limburg_1851 , ontolex:lex_limburg_1825 , ontolex:lex_limburg_1855 ,
ontolex:lex_limburg_1838 , ontolex:lex_limburg_1852 , ontolex:lex_limburg_1856 ,
ontolex:lex_limburg_733 , ontolex:lex_limburg_1837 , ontolex:lex_limburg_1827 ,
ontolex:lex_limburg_608 , ontolex:lex_limburg_5 , ontolex:lex_limburg_1839 ,
ontolex:lex_limburg_1843 , ontolex:lex_limburg_1745 , ontolex:lex_limburg_1842 ,
ontolex:lex_limburg_1823 , ontolex:lex_limburg_204 , ontolex:lex_limburg_1830 ,
ontolex:lex_limburg_1822 , ontolex:lex_limburg_1848 , ontolex:lex_limburg_1835 ,
ontolex:lex_limburg_1836 , ontolex:lex_limburg_1849 , ontolex:lex_limburg_1850 ,
ontolex:lex_limburg_1854 , ontolex:lex_limburg_525 , ontolex:lex_limburg_1817 ,
ontolex:lex_limburg_1821 .
```

In this representation, we can see that the sense "concept_limburg_100" has been "denotated_by" (the reverse property of "denotes") many lexical entries. And this relation is being made explicit in the OntoLex model (and can be quantified), which is also a huge advantage, when compared to the original data.

We have also a total of 4,745 lexical entries, which represent the dialectal variations of the suggested 264 concepts expressed in standard Dutch. An example:

```
ontolex:lex_limburg_1894
        a                ontolex:LexicalEntry ;
        rdfs:label       "staartbot" ;
        ontolex:denotes   ontolex:concept_limburg_103 ;
         ontolex:hasPlace  ontolex:loc_limburg_28 , ontolex:loc_limburg_58 ,
    ontolex:loc_limburg_63 .
```

In this example, we can see that a dialectal word "staartbot" is used for denoting the concept "limburg_103", which is in standard Dutch "stuitbeen"" (*coccyx*). We also get the information about the locations in which this word form is used.

To summarize this exercise: the reader can see how all elements of the original Excel file have been encoded as modules in the OntoLex lexicon for Limburg dialects, and that all instances of such modules are linked to each other using explicit and well defined properties. What is missing in our examples are links to external knowledge resources. This is the topic of the next section.

2.3 Linking to external resources

An issue we would like to consider is the possibility of automatically linking to external resources, those being both of linguistic nature or encyclopedic nature. We do not have an answer to this point for the time being. As a heuristic, while knowing that the Limburg lexical data concerns anatomy, and the reference language is standard Dutch, we can automatically query DBpedia for all entries that have a Dutch word marked with the additional "_(anatomy)" extension, such as for example: http://nl.dbpedia.org/page/Hoofd_(anatomie). However, this might only offer a very specific solution. We will study the algorithm implemented by BabelNet[28] for the automatic cross-linking of language resources in the LOD.

2.4 Quality of the source data

A final point we have to make: In the case of the Limburg lexicon described in this chapter, but also in the case of an automated transformation of two TEI-encoded lexicons of dialectal variants of Arabic into a preliminary version of OntoLex[29], we noticed that in a relevant number of cases some fields of the structured data were not correctly filled by those working on the data. In some cases text was added to the TEI slot "sense", for example "?", or "correct?", and it also occurred that two or more values were included in the slot, instead of introducing a new "sense" slot for every meaning to be encoded.

6. Conclusions

We have been testing the use the OntoLex model, with very few additions, for encoding in the LLOD format the lexicographic resources of some participants of the ENeL Network. The next steps will consist of effectively publishing the results in the Linked Data cloud, after curation of some input data and the clarification of copy-rights issues.

Our current work consists of further automatizing the mapping between the original formats of other ENeL dictionaries and investigating more efficient linking strategies

[28] See http://babelnet.org/

[29] See Declerck et al. (2014b)

to encyclopedic sources. We are also extending our work to the encoding of so-called conceptual records used by lexicographers when carrying out field studies: they interview people in certain regions and ask them how they express certain concepts in their language. We started to use the ConceptSet and LexicalConcept constructs of OntoLex for this task.

We also need to establish clear links to temporal information, which is crucial not only for the encoding of etymology, but also for encoding all kinds of examples and publication dates. There is also a need to link certain lexicographic data to location information.

7. Acknowledgements

The work described in this paper is supported in part by the European Union, by the LIDER project (under Grant No. 610782), by the FREME project (under Grant No. 644771) and by the COST Action IS1305 “ENeL”. Our thanks go especially to the participants of the ENeL COST Action who provided for their data and advices.

8. References

Cimiano, P. & Unger, C. (2014). Multilingualität und Linked Data. In: T. Pellegrini, H. Sack & S. Auer (eds.) *Linked Enterprise Data. Management und Bewirtschaftung vernetzter Unternehmensdaten mit Semantic Web Technologien.* Springer, pp. 153-175.

Declerck, T. & Wandl-Vogt, E. (2014). Cross-linking Austrian dialectal Dictionaries through formalized Meanings. In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress*, pp. 329-343.

Declerck, T., Mörth, K. & Wandl-Vogt, E. (2014b). A SKOS-based Schema for TEI encoded Dictionaries at ICLTT, In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014),* pp. 26-31.

Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J.-P., Cimiano, P. & Navigli, R. (2014). A Multilingual Semantic Network as Linked Data: *lemon*-BabelNet. In C. Chiarcos, J.-P. McCrae, P. Osenova & C. Vertan (eds.) *Proceedings of the 3rd Workshop on Linked Data in Linguistics*, pp. 71-76.

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. *Language Resources and Evaluation,* 46(4), pp. 701-719.

Rehm, G. & Sasaki, F. (2014). Semantische Technologien und Standards für das mehrsprachige Europa. In B. Humm, B. Ege & A. Reibold (eds.) *Corporate Semantic Web.* Springer .

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

Chiarcos, C., McCrae, J.-P., Cimiano, P. & Fellbaum, C. (2013a). Towards open data for linguistics: Lexical Linked Data. In A. Oltramari, P. Vossen, L. Qin & and E. Hovy (eds.) *New Trends of Research in Ontologies and Lexical Resources* Springer, Heidelberg.

Chiarcos, C., Moran, S., Mendes P.-N., Nordhoff, S. & Littauer, R. (2013b). Building a Linked Open Data cloud of linguistic resources: Motivations and developments. In Iryna Gurevych and Jungi Kim (eds.) *The People's Web Meets NLP. Collaboratively Constructed Language Resources.* Springer, Heidelberg.