Towards Multilingual eLexicography by Means of Linked (Open) Data

Thierry Declerck[1], Eveline Wandl-Vogt[2], Simon Krek[3] and Carole Tiberius[4]

[1] DFKI GmbH, Multilingual Technologies Lab,
Saarbrücken, Germany
declerck@dfki.de
[2] ACDH, Austrian Academy of Sciences,
Vienna, Austria
Eveline.Wandl-Vogt@oeaw.ac.at
[3] Jožef Stefan Institute,
Ljubljana, Slovenia
simon.krek@ijs.si
[4] Instituut voor Nederlandse Lexicologie,
Leiden, Netherlands
Carole.Tiberius@inl.nl

Abstract. In this short paper, we document the current state of work consisting in mapping various lexicographic resources onto the OntoLex model, which is an OWL and RDF(s) based representation format. This model has been designed in the context of a W3C Community Group effort for supporting the publication of linguistic data in the Linked (Open) Data cloud. The deployment of OntoLex is currently being tested within the ISCH COST Action IS1305 European Network of e-Lexicography (ENeL), which is adapting to the field of digital lexicography guidelines that have been suggested by the LIDER FP7 Support Action.

Keywords: Multilingual Lexicography, Linguistic Linked Open Data

1. Introduction

The European Network of e-Lexicography (ENeL)[1] is a European COST action that aims at increasing, coordinating and harmonizing European research in the field of e-lexicography and to make authoritative information on the languages of Europe easily accessible.

The working groups of ENel deal with the fact that computers and the availability of the World Wide Web (WWW) have changed the conditions for the production and reception of dictionaries. For editors of scientific dictionaries, the WWW is not only a source of inspiration, but also a new and challenging possibility, for example, when it comes to closing the gap between the public and scientific dictionaries, while ensuring users easy access to scientific dictionaries. ENeL also attempts to provide a

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

broader and more systematic exchange of know-how and common standards and solutions in the field of lexicography. In addition, the pan-European nature of the lexicographical work in Europe is central to ENeL.

This effort involves the exchange of resources, technologies and experience in e-lexicography and provides support for dictionaries which are not yet online. A focal point of ENeL consists in discussing and establishing standards for innovative e-dictionaries that fully exploit the possibilities of the digital medium. In doing so, ENeL explores new ways of representing the common heritage of European languages by developing shared editorial practices and by interconnecting already existing information.

For participants of the Working Group 3 "Innovative eDictionnaries" of ENeL it rapidly seemed obvious that the expanding Linked Open Data (LOD) framework[2], and more specifically the emerging Linguistic Linked Open Data (LLOD)[3], could offer a potential infrastructure for realizing some of its goals. In the next sections we shortly present the main principles of the LLOD and its core representation format, the Ontolex model[4], before describing the current state of our work in mapping various lexicographic resources of the ENeL participants to the Ontolex model.

2. Linguistic Linked (Open) Data

Wikipedia gives the following definition of Linked Data: "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"[5]. Data sets that have been published in the linked data format can be visualized by the so-called Linked Open Data Cloud diagram[6] or also by other representations like the Linked Open Data Graph[7].

In the context of this further expanding Linked Data framework, work has started in encoding linguistic resources in the same format as already existing linked data sets, which were primarily consisting 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.

Recently, some researchers 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 this is the case for the encyclopedic knowledge they were

[2] See http://linkeddata.org/ for more details.
[3] See http://linguistics.okfn.org/resources/llod/ for more details.
[4] See https://www.w3.org/community/ontolex/ for more details.
[5] http://en.wikipedia.org/wiki/Linked_data
[6] http://lod-cloud.net/
[7] http://inkdroid.org/lod-graph/

"commenting" and "labeling"[8]. Cooperation on those topics has been established between, among others, the Working Group on Open Data in Linguistics[9] and the European FP7 Support Action "LIDER"[10]. Those joined efforts have led to the establishment of a Linked Open Data (sub-)cloud of linguistic resources, which is called Linguistic Linked Open Data (LLOD)[11]. The Linguistic Linked Open Data cloud is also visualized by an on-line diagram[12], which itself is derived from information contained in the LingHub[13] repository developed in the context of the LIDER project.

At the core of the publication of language data and linguistic information in the LLOD there is the model "Ontolex" resulting from the W3C Ontology-Lexicon community group[14]. Since this model was originally based on LMF, 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.

3. OntoLex

The OntoLex model is based on the ISO Lexical Markup Framework (LMF)[15] and is an extension of the *lemon* model, which is described in [5]. Ontolex describes a modular approach to lexicon specification, allowing thus the eLexicographer 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 now is 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 relation markers. Now, the components of a dictionary entry can be distributed in a network and be 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 the Web of data.

Practically, this means that a dictionary author does not need to describe all components or elements of an entry in details, 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

[8] See for example [8] and [9].
[9] http://linguistics.okfn.org/.
[10] See http://lider-project.eu/ for more details.
[11] See http://linguistics.okfn.org/tag/llod/ for more details.
[12] http://linguistic-lod.org/llod-cloud
[13] See http://linghub.lider-project.eu/about for more details.
[14] See http://www.w3.org/community/ontolex/ for more details.
[15] See [7] and http://www.lexicalmarkupframework.org

between various scientific lexicographers, and that this can result in virtual and collaborative research environments in the lexicographical field.

Fig. 1 below displays the core model of Ontolex[16]. 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.

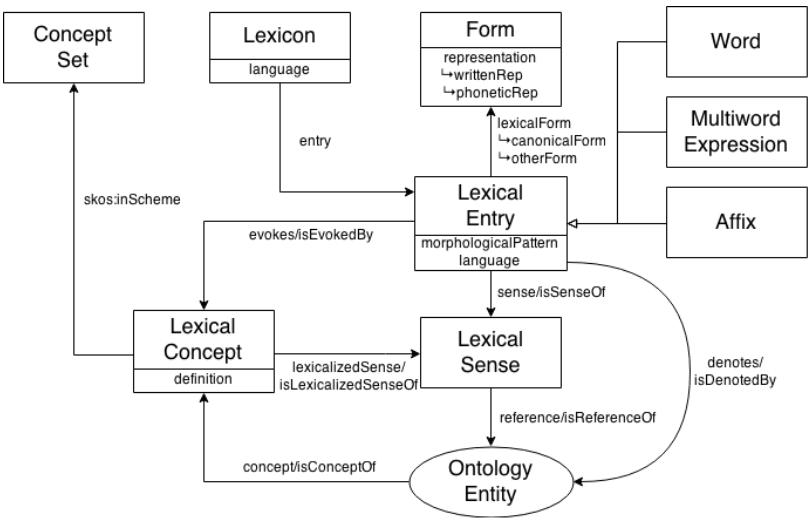


Fig. 1. The core model of Ontolex. Figure created by John P. McCrae for the W3C Ontolex Community Group.

We used this model on a list of lexical resources made available by participants of the ENeL network, and we describe this transformation process in the next section.

4. Mapping ENeL lexicographic Resources onto Ontolex

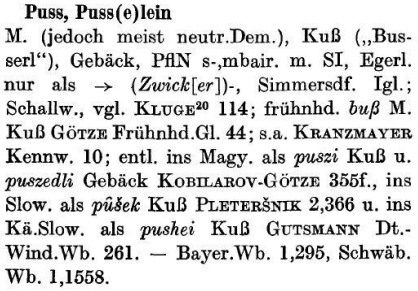
In order to test our intuition about the format for the publication of lexicographic resources in the LOD, cooperation between the ENeL COST action and the LIDER projects has therefore been established. We got from ENeL participants, for a first test, samples from 13 dictionaries, which are in different languages:

- 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)

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

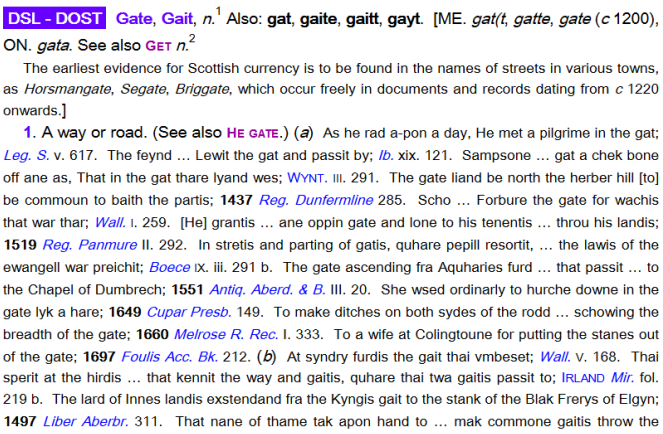
- 1 Sample from a French lexicon (extracted from Wiktionary)
- 1 Limburg lexicon (Excel)
- 1 Sample from the KDictionary multilingual source (XML file)
- 1 Sample from the Digital Scottisch Lexicon (Old Scottisch, html + 1 example in TEI)
- 1 Lexicon extracted from a corpus of „Baroque German“

Below in figures 2 and 3, the reader can see the kind of data we are dealing with. In Figure 2, we have an example taken from a longer entry of the Austrian Dictionary of Bavarian Dialects[17] and in Figure 3 a screen shot from the web page of the Dictionary of the Older Scottish Tongue[18]



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.

Fig. 2. An example entry from the Austrian Bavarian Dictionary (without the usage examples)



DSL - DOST **Gate, Gait**, *n.*[1] Also: **gat, gaite, gaitt, gayt**. [ME. *gat(t, gatte, gate* (*c* 1200), ON. *gata*. See also GET *n.*[2]

The earliest evidence for Scottish currency is to be found in the names of streets in various towns, as *Horsmangate, Segate, Briggate*, which occur freely in documents and records dating from *c* 1220 onwards.]

1. A way or road. (See also HE GATE.) (*a*) As he rad a-pon a day, He met a pilgrime in the gat; *Leg. S.* v. 617. The feynd ... Lewit the gat and passit by; *Ib.* xix. 121. Sampsone ... gat a chek bone off ane as, That in the gat thare lyand wes; WYNT. III. 291. The gate liand be north the herber hill [to] be commoun to baith the partis; **1437** *Reg. Dunfermline* 285. Scho ... Forbure the gate for wachis that war thar; *Wall.* I. 259. [He] grantis ... ane oppin gate and lone to his tenentis ... throu his landis; **1519** *Reg. Panmure* II. 292. In stretis and parting of gatis, quhare pepill resortit, ... the lawis of the ewangell war preichit; *Boece* IX. iii. 291 b. The gate ascending fra Aquharles furd ... that passit ... to the Chapel of Dumbrech; **1551** *Antiq. Aberd. & B.* III. 20. She wsed ordinarly to hurche downe in the gate lyk a hare; **1649** *Cupar Presb.* 149. To make ditches on both sydes of the rodd ... schowing the breadth of the gate; **1660** *Melrose R. Rec.* I. 333. To a wife at Colingtoune for putting the stanes out of the gate; **1697** *Foulis Acc. Bk.* 212. (*b*) At syndry furdis the gait thai vmbeset; *Wall.* v. 168. Thai sperit at the hirdis ... that kennit the way and gaitis, quhare thai twa gaitis passit to; IRLAND *Mir.* fol. 219 b. The lard of Innes landis exstendand fra the Kyngis gait to the stank of the Blak Frerys of Elgyn; **1497** *Liber Aberbr.* 311. That nane of thame tak apon hand to ... mak commone gaitis throw the

Fig. 3. An example from the Dictionary of the Older Scottish Tongue

In Figure 2 we can observe a property of many dialectal or regional dictionaries. They express the meaning of the entries by using words taken from the standard language. The meaning of “Puss” or “Puss(e)lein” is expressed by the standard

[17] See http://verlag.oeaw.ac.at/Woerterbuch-der-bairischen-Mundarten-in-Oesterreich-38.-Lieferung-WBOe for more details.
[18] See http://www.dsl.ac.uk/ for more details.

German word “Kuss” (*kiss*), or “Gebäck” (*pastry*). The third meaning of the entry is expressed by an abbreviation expressing a category “PflN” (meaning *name of a plant*). We find in this (part of the) entry also etymological information (“frühnhd.”, this being the abbreviation for *Early New High German*). The entry is marked as being a masculine word, but used mostly in neutrum. A Bavarian variant of the word with the *kiss* meaning is also given *(“Busselr”).* And much more (a long list of examples of usages is also given in the full entry, with some additional definition text) is available. It is important to note here, that in the digital version of the dictionary we have at our disposal, and which is encoded in the TUSTEP[19] format for supporting publication, we had to gather much of the information ourselves. The metadata information is very poor. This kind of dictionary was in the past in fact more directed to the professional lexicographer rather than the general public, and many interpretation aspects of the codes used for the entry were supposed to be known by the reader. So for example the typographical coding of the headword (what we call here “entry”) can include some information which we had to gain from an annex or directly from the lexicographers.

Similar comments apply to the example in Figure 3. There we had for example to infer the interpretation of the different temporal expressions (the TEI code of the dictionary has been provided to us as a sample only for one entry). And we also had to interpret the typographical codes used.

Therefore a manual analysis of the resources we got from the ENeL participants was needed in order to know if and how an automatized mapping to Ontolex can be implemented. Also we had to add some few classes and properties to the Ontolex model in order to deal with certain features of the dictionaries. For example we added a class for the etymology, a class for describing the lexicographic slips used by lexicographer and some properties to encode the different types of temporal information (date of publication vs etymological information etc.). For most of the lexical information encoded in the 13 dictionaries, we could find a way to map it to the Ontolex model (see Fig. 1). Every dictionary has been encoded as an ontolex:lexicon, using the ontolex:entry object property to indicate inclusion of an entry:

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

The entries are instances of the ontolex:LexicalEntry class and ambiguities are marked by introducing various instances of the ontolex:LexicalSense class.

```
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 ;
```

[19] See http://www.tustep.uni-tuebingen.de/tustep_eng.html for more details.

```
  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 ;
.

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> ;
.
```

The use of the properties "`ontolex:sense`" and "`ontolex:denotes`" is very important if one wants to link lexical resources in a multilingual way, just looking if they are sharing the same senses. The difference between the two properties is that the first one is pointing to instances of the class "LexicalSense", which is collecting ontological objects within the model, while the second property points directly to external resources. Instances of the class "LexicalSense" are linked to external knowledge resources via the property `ontolex:reference`. Figure 1 in chapter 3 above is graphically representing the difference between the usages of the two properties "reference" and "denotes".

On the basis of the use of the Ontolex model we could semi-automatically establish not only links between entries within and between the samples of the ENeL dictionaries, but also links to encyclopedic data sets in the LOD, like for example DBpedia[20] or the BabelNet resource[21], which is automatically merging various multilingual language and encyclopedic resources that are available in RDF.

BabelNet is in fact an excellent example of such a combination of linguistic and encyclopedic data in the LOD cloud. All language data are encoded in RDF and *lemon* (the former version of OntoLex). While BabelNet was considering mainly the RDF Version of WordNet and collaboratively created lexical resources, like Wiktionary, our work is aiming at adding to this framework the language and encyclopedic data that has been created and published by professional lexicographers.

5. Conclusions and future Work

We could successfully use the Ontolex model, with very few additions, for encoding in the LLOD format the lexicographic resources of some participants of the ENeL Network. Next steps will consist in effectively publish the results in the Web.

Our current work consists in further automatizing the mapping between the original formats of other ENeL dictionaries and in investigating more efficient linking strategies to encyclopedic sources. We are also extending our work to the encoding of so-called conceptual records used by lexicographers when doing field studies: the

[20] http://dbpedia.org/

[21] http://babelnet.org/

interview people in certain regions and ask them how they express in their language certain concepts. We started to use the ConceptSet and LexicalConcept constructs of Ontolex for this task.

Acknowledgments

The work described in this short paper submission is supported in part by the European Union, both by the LIDER project (under Grant No. 610782) and by the COST Action IS1305 "ENeL". Our thanks go also to the participants of the ENeL COST Action, who provided for their data and advices. And finally our thanks go to the anonymous reviewers of the first version of this paper, helping a lot to improve it.

References

1. Thierry Declerck, Eveline Wandl-Vogt. Cross-linking Austrian dialectal Dictionaries through formalized Meanings. In: Andrea Abel, Chiara Vettori, Natascia Ralli (eds.): Proceedings of the XVI EURALEX International Congress, Pages 329-343 (2014)
2. Thierry Declerck, Eveline Wandl-Vogt. How to semantically relate dialectal Dictionaries in the Linked Data Framework. Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2014), Gothenburg, Sweden, ACL (2014)
3. Maud Ehrmann, Francesca Cecconi, Daniele Vannella, John P. McCrae, Philipp Cimiano, and Roberto Navigli. A Multilingual Semantic Network as Linked Data: *lemon*-BabelNet. Proceedings of the 3rd Workshop on Linked Data in Linguistics (2014)
4. Philipp Cimiano and Christina Unger. Multilingualität und Linked Data. In: Tassilo Pellegrini, Harald Sack, and Sören Auer (eds): Linked Enterprise Data. Management und Bewirtschaftung vernetzter Unternehmensdaten mit Semantic Web Technologien (2014)
5. J. McCrae, G. Aguado-de-Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, T. Wunner. Interchanging lexical resources on the Semantic Web. Language Resources and Evaluation (2012).
6. Georg Rehm and Felix Sasaki. Semantische Technologien und Standards für das mehrsprachige Europa. In: B. Humm Ege, B. and A. Reibold (eds.): Corporate Semantic Web (2014)
7. Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, Claudia Soria. Lexical Markup Framework (LMF). In: Proceedings of the fifth international conference on Language Resources and Evaluation (2006)
8. Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum. Towards open data for linguistics: Lexical Linked Data. In Alessandro Oltramari, Piek Vossen, Lu Qin, and Eduard Hovy (eds.): New Trends of Research in Ontologies and Lexical Resources Springer, Heidelberg (2013)
9. Christian Chiarcos, Steven Moran, Pablo N. Mendes, Sebastian Nordhoff, and Richard Littauer. 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 (2013)