Encoding lexicographic Data in *lemon*: Lessons learned

Thierry Declerck[1], Carole Tiberius[2], Eveline Wandl-Vogt[3]

[1] DFKI GmbH, Multilingual Technologies Lab
Stuhlsatzenhausweg.3, 66123 Saarbrücken, Germany
declerck@dfki.de
[2] Dutch Language Institute
Postbus 9515, 2300 RA Leiden, Netherlands
Carole.Tiberius@ivdnt.org
[3] Austrian Centre for Digital Humanities,
Austrian Academy of Sciences
Sonnenfelsgasse 191010 Vienna, Austria
eveline.wandl-vogt@oeaw.ac.at

Abstract. We describe experiments done in using the *lemon* model for encoding lexicographic data we got from different sources with distinct coverages. Our focus is on delivering statements on lessons learned and on questions that still should be discussed in the *lemon* community, as a possible input for forthcoming versions of the model.

Keywords: lexicography, *lemon* model

1 Introduction

In recent years, we have been experimenting with the use of the Lexicon Model for Ontologies (*lemon*)[1] for representing lexicographic data. *lemon* has been developed within the W3C Ontology-Lexica community group[2], building on the version that was first proposed in the context of the European R&D project "Monnet"[3]. Members of the W3C community group started to investigate, in close cooperation with members of the ENeL Cost Action[4], on how this model could be used and possibly extended for the purpose of the encoding of rich lexicographic data in the context of the Linguistic Linked Open Data (LLOD) framework[5].

---

1 See https://www.w3.org/2016/05/ontolex/ [accessed 13.07.2017, like all other URLs mentioned in this paper]

2 https://www.w3.org/community/ontolex/

3 http://cordis.europa.eu/project/rcn/93713_en.html. See for the first version of *lemon*: http://lemon-model.net/

4 ENeL stands for "European Network of e-Lexicography". See http://www.elexicography.eu/ for more details on this COST Action.

5 See http://linguistic-lod.org/ for more details.

The studies we present in this paper are dealing with the "Wörterbuch der bairischen Mundarten in Österreich" (WBÖ)[6] and with the "Algemeen Nederlands Woordenboek" (ANW)[7]. Details on the first study are described in [1] and for the second study in [6]. In this paper we focus on lessons learned and on questions that still should be discussed in the W3C Ontology-Lexica community, as a possible input for future versions of the *lemon* model.

2 The *lemon* model

The original aim of *lemon* was to provide for a rich linguistic description for natural language expressions used in knowledge resources, like taxonomies or ontologies. This linguistic grounding includes the formal representation of morphological and syntactic properties of lexical entries as well as the specification of the meaning of these lexical entries with respect to available knowledge resources, more specifically formal ontologies. This relation between lexical entries and their meaning to be found in external ontologies is specified in a core module, which is realising the so-called ontology-lexicon interface (ontolex). Figure 1 below is giving a graphical overview of the ontolex module.

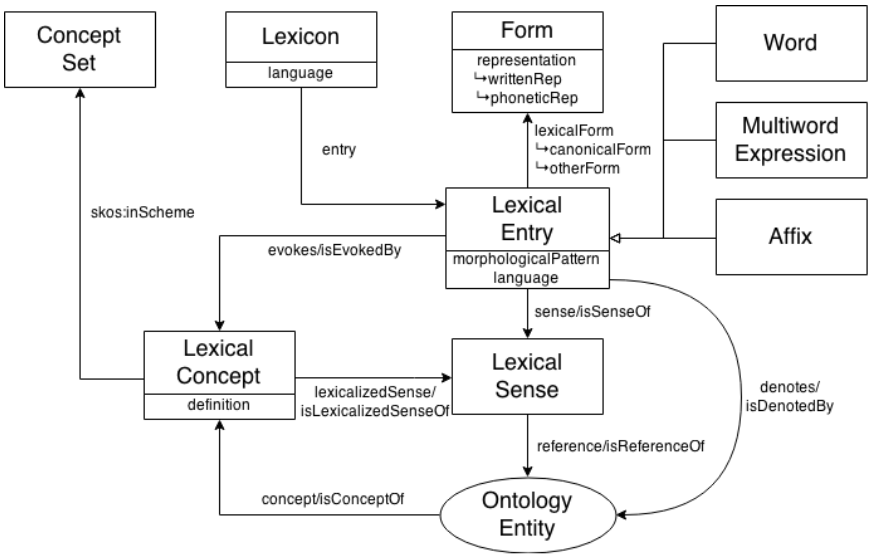

Fig. 1. Ontolex: the core module of *lemon*: Figure created by John P. McCrae for the W3C Ontology-Lexica Community Group

[6] Dictionary of the Bavarian Dialects in Austria: https://www.oeaw.ac.at/acdh/projects/wboe/.
[7] General Dutch Dictionary: http://anw.inl.nl/

The full list of *lemon* modules comprises the following items[8]:

- Ontology-lexicon interface (ontolex), which is the core module of *lemon*
- Syntax and Semantics (synsem)
- Decomposition (decomp)
- Variation and Translation (vartrans)
- Linguistic Metadata (lime)

3 The "Wörterbuch der bairischen Mundarten in Österreich"

Our first experiment dealing with the encoding of lexicographic data in *lemon* involved the "Wörterbuch der bairischen Mundarten in Österreich" (WBÖ)[9], which is a large-scale dictionary documenting (spoken) languages used in Austria and neighboring areas such as South Tyrol. WBÖ describes a very large varietal spectrum of the language spoken in the regions it covers. In addition to "Austriacisms", the publication also contains standard German keywords, which have developed a different range of meaning in the dialects.

3.1 The Dictionary Data of WBÖ

The basis we are working on is an XML representation of the electronic version of the WBÖ dictionary. As the WBÖ is primarily aimed at scientists it uses for its descriptors complex and rich representation forms. So for example, the strings for representing headwords in WBÖ can include information about pronunciation or word formation properties[10], etc. A first issue for the encoding of the WBÖ headwords in *lemon-ontolex* concerns thus the status of this highly specialized and compact representation form. As the *lemon* model considers usually the "lemma form" of a lexical entry as its canonical form[11], we would tend to give to the corresponding instance of the `ontolex:Form` class, as the range of the associated object property `ontolex:canonicalForm`, the lemma form of the original headword as the value of datatype property `ontolex:writtenRep`, without the special characters used with the original headword. And we would encode for example the pronunciation properties as a value of the related datatype property "`ontolex:phoneticRep`"[12]. Thus marking explicitly all the information related to a headword by corresponding properties.

The issue we have is how to encode the original headword. We do not consider it as being potentially an "other form", in the sense this feature has been introduced in *lemon-ontolex* (and similar also in other vocabularies, like SKOS-XL with the "alter-

[8] Taken from https://www.w3.org/2016/05/ontolex/

[9] Dictionary of the Bavarian Dialects in Austria: https://www.oeaw.ac.at/acdh/projects/wboe/

[10] An example of such a headword is "(Ge)pâcht", where the parentheses mark a prefix and the circumflex refers to a pronunciation property.

[11] See also http://www.w3.org/ns/lemon/ontolex#canonicalForm

[12] See Figure 1 for the roles played by those properties.

native label"). We can encode the original headword as a "hidden form", but there would be a need to associate a comment to this representation, stating that this was the original form used in the dictionary. In general, the issue we are dealing with concerns the wish to keep information about the way the data was represented in the original source. In case we can decompose all the lexical information included in the rich and compact original representation of the headword and port it to different elements of *lemon*, we might then just add a property referring to the original headword string.

We used only the ontolex module for the *lemon* encoding of WBÖ, and as described in [2], it was quite straightforward to encode all the senses related to a headword in WBÖ. One addition that was needed to ontolex is related to etymological information included in WBÖ. For this purpose we need to consider temporal relations that are more detailed than the currently used "outdated" or similar values that are available in the LexInfo vocabulary. And WBÖ being a dictionary about dialectal variations, information about locations are also needed.

We see in the current specifications of the "vartrans" module[13] that it covers among others both diatopic and diachronic lexical variations, but the specifications also encourage the use of external lexical vocabularies for marking temporal information on the usage of a word, this being also relevant for synchronic lexicography. It will be important to reach a consensus on which vocabulary to use for indicating temporal and local information, similar to the use of the LexInfo vocabulary[14] for marking many morphological and syntactic properties of lexical entries within ontolex.

3.2 The "Questionnaires'' used for the WBÖ Creation

A very interesting resource associated with WBÖ is provided by the ("Fragebögen") (questionnaires) that were used for interviewing people in different regions of Austria on which words they use for expressing a specific concept. The questionnaires included concepts and related definitions that could also represent various senses associated with one term. Those questionnaires form an important part of the very rich documentary material that was used for the creation of the WBÖ, and which contains an estimated 4 million individual sheets[15]. This documentary material was entered and digitized between 1993 and 2011 in the so-called database of the Bavarian dialects in Austria (DBÖ)[16].

As those questionnaires represent a very rich combination of conceptual and related lexical information, we also investigated the possibility of encoding this information in *lemon-ontolex*. It turned out that the possibility to link lexical entries (and also lexical senses) to a lexical concept, which is itself a member of a SKOS

---

[13] See for more details on this module: https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans

[14] See http://www.lexinfo.net/ for more details.

[15] This information is taken from the German Wikipedia page: https://de.wikipedia.org/wiki/W%C3%B6rterbuch_der_bairischen_Mundarten_in_%C3%96sterreich

[16] See https://wboe.oeaw.ac.at/projekt/beschreibung/

scheme[17], is very suitable to represent the type of information that is encoded in both the questionnaires and the WBÖ dictionary. The questionnaire has been encoded in the SKOS vocabulary and the ontolex property "isEvokedBy" is linking the concepts of the questionnaires to the *lemon-ontolex* lexical entries, while the ontolex property "lexicalizedSense" is linking the concepts to corresponding WBÖ lexical senses. It is unclear, if we can still speak of a mental abstraction or unit of thought of lexical entries (as the introduction of the "LexicalConcept" class in ontolex was aiming at), but the fact is that we can efficiently relate the conceptual background that was developed as a basis for the creation of a dialectal dictionary to the Lexical Entries in the *lemon* representation of this dictionary.

Based on our piloting work, the questionnaires are now completely conceptually interlinked and all sources made available for conceptual based discovery within the project exploreAT! [18]

3.3 The Paper Slips used for the WBÖ Creation

An additional artefact that was used for crafting the dictionary is a (huge) set of paper slips, on which the field lexicographer was indicating the answer of the interviewed persons, with some metadata (location, time, and any other comments). Those paper slips are for sure an important element of cultural heritage, but besides this they also offer – together with the questionnaires – a view on the methodology, the "workflow" and the material used. We do not foresee to encode this data in *lemon-ontolex*. We rather propose to build for this combination of the two artefacts, questionnaires and paper slips, a model using SKOS, and linking the concepts of those SKOS schemes to the lexical entries we have in *lemon-ontolex*. All the dialectal variants that were encoded in the paper slips (and those already present in WBÖ) will then be encoded in the *lemon* module "vartrans".

3.4 First Conclusions

Our aim in the WBÖ case was to develop a series of methodological prototypes of a machine-readable and modular version of the lexicographical work, aiming at making it available in the Linguistic Linked Open Data framework. We also suggest ways for encoding in SKOS some information on data material that was used for the creation of the original dictionary. But for this, the open question remains if we should aim at creating a new module in *lemon*, which is collecting all those aspects of a lexicographic work, and not only the lexical knowledge expressed in the dictionary.

Our learnings and prototypes are further explored within the project exploreAT! to open up the resources for multicultural, multilingual knowledge discovery.

---

[17] See again for details Figure 1.

[18] exploreAT! exploring austria´s culture with the language glass. https://exploreat.grial.eu/dashboard (last accessed: June, 12th 2017).

4 The „Algemeen Nederlands Woordenboek“

A second source of data we have been working with is the XML representation of the online version of the Algemeen Nederlands Woordenboek (ANW). ANW is an online, corpus-based, scholarly dictionary of contemporary standard Dutch in the Netherlands and in Flanders, describing the Dutch vocabulary from 1970 onwards (see [3]). One of the innovative features of the ANW is that it offers a twofold meaning description: definitions are accompanied by a semagram, a frame-based representation of knowledge typically associated with a word (see [4]).

The ANW is a digitally-born dictionary. It is very rich data encoded in a format that is already very abstract. With the addition of the semagram framework, the ANW includes also some accompanying data structure reflecting the conceptual world to which lexical entries are related. A nice aspect of the ANW is that it also refers to an external resource for the information on etymology. Thus a high level of modularity and connectivity is already realized.

Like a number of scholarly dictionaries, ANW has a large number of senses associated to the entries (at least for certain categories of entries – nouns, verbs, etc.). As this repository of senses is large and complex, a discussion arose if we should not introduce in *lemon* a sub-sense hierarchy, which is present in ANW, where a numbering strategy is used for representing the hierarchy of senses. We note that a former version of the *lemon* model[19] included the notion of “subsense”. But the property defined in this version was meant to describe the composition of senses needed for describing the composition of senses resulting from the argument structure of a verb (or another category introducing syntactic arguments). In the new version of *lemon*, resulting from the W3C community group discussions, this aspect is dealt with by the synsem module and the notion of “subsense” has disappeared.

For the time-being we suggested to have in *lemon-ontolex* a flat list of senses and to interrelate those by the use of corresponding properties, such as `lexinfo:hypernym`, `lexinfo:partHolonym`, `lexinfo:substanceMeronym`, `lexinfo:pertainsTo`, etc. As a matter of fact, it turned out that the LexInfo vocabulary is here (and in general) very helpful for describing relational properties of lexical senses. But we kept the original ANW numbering for naming the senses object in *lemon-ontolex*. Extending the study we describe in this paper to other scholarly dictionaries will certainly help in getting final decisions on this issue.

For the porting of the verbs included in the ANW we also started to investigate the use of the synsem module. A first comment would be that we are not sure about the necessity to use the ontology mappings, as described in the specification document[20]. It seems to introduce a higher level of complexity. And the need to introduce a frame element for each verb seems also to introduce a lot of redundancies. But we have no alternative solutions for the time being. Our tentative modelling of the verb “eten” (*to eat*) looks like in the following:

---

[19] http://lemon-model.net/lemon-cookbook.pdf

[20] See again https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

```
:lex_eten_47968
  rdf:type ontolex:Word ;
  lexinfo:partOfSpeech lexinfo:verb ;
  ontolex:canonicalForm :form_eten_infinitive ;
  ontolex:sense <http://tutorial-topbraid.com/anw#sense_eten0.1>
;
  synsem:synBehavior :eten_frame_1 ;
.

<http://tutorial-topbraid.com/anw#sense_eten0.1>
  rdf:type ontolex:LexicalSense ;
  skos:definition "iets als voedsel tot zich nemen; iets opeten;
iets nuttigen"@nl ; #(take something as food, eat something, …)
  ontolex:isLexicalizedSenseOf :Semagram_activiteit ;
  ontolex:isLexicalizedSenseOf :Semagram_handeling ;
  ontolex:isSenseOf :lex_eten_47968 ;
  synsem:objOfProp :eten_frame_obj_1 ;
  synsem:subjOfProp :eten_frame_subj_1 ;
.

:eten_frame_1
  rdf:type lexinfo:TransitiveFrame ;
  rdf:type synsem:SyntacticFrame ;
  lexinfo:directObject :eten_frame_obj_1 ;
  lexinfo:subject :eten_frame_subj_1 ;
  rdfs:comment "one syntactic frame for the Dutch verb
\"eten\""@en ;
  rdfs:label "transitief eten"@nl ;
.

:eten_frame_subj_1
  rdf:type lexinfo:Subject ;
  rdf:type synsem:SyntacticArgument ;
  rdfs:comment "A subject of the eten_frame" ;
  rdfs:label "subject 1 for eten_frame"@en ;
  ontolex:concept :SemaGram_dier ;
  ontolex:concept :SemaGram_mens ;
.

:eten_frame_obj_1
  rdf:type lexinfo:DirectObject ;
  rdf:type synsem:SyntacticArgument ;
  rdfs:comment "An object of the eten_frame"@en ;
  rdfs:label "object 1 for eten_frame"@en ;
  ontolex:concept :SemaGram_voedsel ;.
```

```
:OntoMap_eten_1
  rdf:type synsem:OntoMap ;
  rdfs:comment "Mapping the syntacic frame eten_1 with seman-
tics" ;
  rdfs:label "OntoMap_eten_1@en}" ;
  synsem:objOfProp :eten_frame_obj_1 ;
  synsem:ontoMapping <http://tutorial-
topbraid.com/anw#sense_eten0.1> ;
  synsem:subjOfProp :eten_frame_subj_1 ;
.
```

But as mentioned above, we have the feeling that this representation is getting too complex.

We also tested the decomposition module for encoding Dutch compounds with the example word being “wijnfles” (*bottle of wine*). As above for the “eten” verb, we just display the current code, so that the reader can get a concrete idea of the possibility offered by this “decomposition” module of *lemon*:

```
:wijnfles
  rdf:type ontolex:MultiWordExpression ;
  <http://lemon-model.net/lexinfo_anw#articleType> "\"de\"" ;
  lexinfo:gender lexinfo: commonGender ;
  lexinfo:partOfSpeech lexinfo:commonNoun ;
  lexinfo:partOfSpeech lexinfo:noun ;
  rdf:_1 :comp_wijn_1 ;
  rdf:_2 :comp_fles_1 ;
  <http://www.w3.org/ns/lemon/decomp#constituent> :comp_fles_1 ;
  <http://www.w3.org/ns/lemon/decomp#constituent> :comp_wijn_1 ;
  <http://www.w3.org/ns/lemon/decomp#subterm>
<http://dictionary_lemon/anw#lex_wijn_182155> ;
  <http://www.w3.org/ns/lemon/decomp#subterm> :lex_fles_18089 ;
  ontolex:sense <http://tutorial-topbraid.com/anw#sense_wijn1.3>
;
.

:lex_wijn_182155
  rdf:type ontolex:Word ;
  <http://lemon-model.net/lexinfo_anw#articleType> "\"de\"" ;
  lexinfo:gender lexinfo:masculine ;
  lexinfo:partOfSpeech lexinfo:commonNoun ;
  lexinfo:partOfSpeech lexinfo:noun ;
  ontolex:canonicalForm :form_wijn_singular ;
  ontolex:otherForm :form_wijnen_plural ;
```

```
  ontolex:sense <http://tutorial-topbraid.com/anw-
entry#sense_wijn1.0> ;
.

:lex_fles_18089
  rdf:type ontolex:Word ;
  <http://lemon-model.net/lexinfo_anw#articleType> "\"de\"" ;
  lexinfo:gender lexinfo:feminine ;
  lexinfo:gender lexinfo:masculine ;
  lexinfo:partOfSpeech lexinfo:commonNoun ;
  lexinfo:partOfSpeech lexinfo:noun ;
  ontolex:canonicalForm :form_fles_singular ;
  ontolex:otherForm :form_flessen_plural ;
.

:comp_fles_1
  rdf:type <http://www.w3.org/ns/lemon/decomp#Component> ;
  <http://www.w3.org/ns/lemon/decomp#correspondsTo>
<http://dictionary_lemon/anw#lex_wijn_182155> ;
  <http://www.w3.org/ns/lemon/decomp#correspondsTo>
:lex_fles_18089 ;
.
:comp_wijn_1
  rdf:type <http://www.w3.org/ns/lemon/decomp#Component> ;
  <http://www.w3.org/ns/lemon/decomp#correspondsTo>
<http://dictionary_lemon/anw#lex_wijn_182155> ;
  <http://www.w3.org/ns/lemon/decomp#correspondsTo>
<http://tutorial-topbraid.com/anw#sense_wijn1.0> ;
.
```

The only addition we suggest here, is to add the possibility to have a sense as the value of the “correspondsTo” property, as this can be seen for the element “comp_wijn_1” above. The point being that in this case the component “wijn” can only refer to the generic use of the word, which is the one covered by the Lexical Sense “sense_wijn1.0”, as can be seen below:

```
<http://tutorial-topbraid.com/anw#sense_wijn1.0>
  rdf:type ontolex:LexicalSense ;
  skos:definition "alcoholhoudende drank, verkregen door gisting
van het sap van druiven of van andere vruchten, met een middel-
matig alcoholgehalte van doorgaans ongeveer 12 procent; alco-
holhoudende drank van gegist druivensap" ;
  ontolex:isLexicalizedSenseOf :Concept_325624 ;
  ontolex:isLexicalizedSenseOf :Concept_Stofnaam ;
  ontolex:isSenseOf :lex_wijn_182155 ;
```

```
  ontolex:reference <https://www.wikidata.org/wiki/Q282> ;
  ontolex:usage lexinfo:massNoun ;
  ontolex:usage lexinfo:singular ;
.
```

This way, it seems that we can cover most (if not all) of the relevant conceptual and lexical elements included in the ANW offer. A remaining question being if one needs to introduce in *lemon* a hierarchy of senses, or rather, like we opted for now, if one can adopt a single listing of Lexical Senses and to explicitly mark the relation among them by the use of a possibly extended LexInfo vocabulary.

5 Conclusions

It is our conviction that a lot of the original dictionary data, in different formats and with different coverages, can be accurately modelled with the *lemon* modules. Relevant lexicographic information that is not directly related to the description of the entries (in the sense of providing knowledge about the words) can be designed in or re-used from models external to *lemon*, but a consensus building on the best vocabularies to be used will be needed in this case.

And our current intuition is that elements in *lemon* should not include (deeper) hierarchical structures but represent the relation between elements of the lexicon by the use of specialized properties. We think this is an aspect that should be discussed and possibly fixed within the W3C Ontology-Lexica community.

References

1. Declerck, T., Wandl-Vogt, E.: Cross-linking Austrian dialectal Dictionaries through formalized Meanings. In: Abel, A., Vettori, C., Ralli, N. (eds.) *Proceedings of the XVI EURALEX International Congress*, pp. 329-343, EURAC research, Bolzano/Bozen (2014)
2. Declerck, T., Wandl-Vogt, E.: How to semantically relate dialectal Dictionaries in the Linked Data Framework. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (LaTeCH 2014), Gothenburg, Sweden (2014)
3. Moerdijk, F.; Frames and Semagrams. Meaning Description in the General Dutch Dictionary. In: Berndal, E., De Cesaris, J. (eds) *Proceedings of the XIII EURALEX International Congress*, Barcelona (2008)
4. Schoonheim, T., and Tempelaars, R.: Dutch Lexicography in Progress, The Algemeen Nederlands Woordenboek (ANW). In: Dykstra, A., Schoonheim, T. (eds) *Proceedings of the XIV Euralex International Congress*. Leeuwarden (2010)
5. Cimiano, P., McCrae, John P. and Buitelaar, P. (eds) Lexicon Model for Ontologies: Community Report, 10 May 2016 (2016)
6. Tiberius, C. and Declerck, T.: A lemon model for the ANW dictionary. In *Proceedings of the fifth biennial conference on electronic lexicography*, eLex 2017, Leiden (2017)
7. Datenbank der bairischen Mundarten in Österreich electronically mapped. https://wboe.oeaw.ac.at/projekt/beschreibung/