Towards a Linked Lexical Data Cloud based on OntoLex-Lemon

Thierry Declerck[1,2]
[1]DFKI GmbH, Multilingual Technologies
Stuhlsatzenhausweg 3, D-66123, Saarbrücken, Germany
[2]Austrian Academy of Sciences, Austrian Centre for Digital Humanities (ACDH)
Sonnenfelsgasse 19, A-1010 Vienna, Austria
declerck@dfki.de

Abstract

In this paper, we present some considerations on the current state of both the Linguistic Linked Open Data (LLOD) cloud and the core module of the OntoLex-Lemon model. It is our perception that the LLOD is lacking a representation and interlinking of lexical data outside of the context of lexicons or dictionaries, which have been ported to Linked Data compliant formats. And while the OntoLex-Lemon model and its predecessor *lemon* have originally been developed to support the formal representation of language data used in ontologies, the models have been increasingly used for representing lexical entries of dictionaries and lexicons, as this can be seen in corresponding data sets included in the LLOD. As a consequence of that, we are proposing slight modifications of the core module of OntoLex-Lemon, its ontology-lexicon interface, in order to support the representation and linking of lexical data that are not necessarily included in a lexicon, a dictionary or in the terminology used in a knowledge base.

Keywords: lexical data, Linguistic Linked Open Data, OntoLex-Lemon

1. Introduction

The rapid development of the Linguistic Linked Open Data (LLOD) cloud[1] is a success story that is also based on the development of the Lexicon Model for Ontologies (*lemon*)[2] and its successor, OntoLex-Lemon[3], and experience has shown that *lemon* or OntoLex-Lemon can indeed be used for a variety of applications that are not explicitly related to ontologies, like the modelling of lexicographic data[4] or specific lexical phenomena[5].

As the possibility to develop new modules for OntoLex-Lemon is currently under discussion[6], certain aspects dealing with its core module, the "ontology-lexicon interface", seem to require some clarifications and adaptations. In this paper, we present some slight modifications to the core module of OntoLex-Lemon in order to support the deployment of a Linked Lexical Data cloud.

The suggestions we present in this context are also influenced and guided by (Gracia et al., 2017), in whose abstract we can read: "[...] future dictionaries could be LD-native and, as such, graph-based. Their nodes do not depend on any internal hierarchy and are uniquely identified at a Web scale". We can clearly see how OntoLex- Lemon is at the core of such a development, not only in the context of LD-native dictionaries, but also for Linked (stand-alone) Lexical Data.

At the same time, we are perfectly aware of the fact that *lemon*, which stands for "LExicon Model for ONtologies", was originally developed in order to model language data used in ontologies[7]. In this original context, our interpretation of "lexicon" describes the collection of language data that are included in labels or comments in ontologies, aiming to give a human-readable description of the knowledge source's content. For modelling this particular language data *lemon* is using the same formal representation language as the one deployed for the knowledge objects they describe. This approach is ultimately supporting the bridging of the knowledge of the world (or of a domain) and the knowledge of the words that are used in the same ontological environment.

However, it rapidly turned out that *lemon* and its successor, the OntoLex-Lemon model, are being used more and more for modelling digital (versions of) lexicons or dictionaries per se[8]. While this constitutes to a highly positive development, we think that a Linked Data (LD)-based lexicographic network could be independent of specific dictionaries or lexicons containing the lexical data to be represented. Quoting again from (Gracia et al., 2017): In a native LD environment "every lexical element (headword, sense, written form, grammatical attribute, etc.) is treated as a first-class citizen, being identified by its own URI at a Web scale, and being attached to its own descriptive information and linked to other relevant elements through RDF statements". While the authors still anchor this view in the context of developing an "LD-based dictionary", we argue that specific dictionaries or lexicons are not necessary as container for representing lexical data in a Linked Data environment.

We consider OntoLex-Lemon as an excellent basis for reaching this goal of a Linked Lexical Data cloud, and in the next sections, we will suggest some slight modifications

[1] `http://linguistic-lod.org/llod-cloud`. See also (Chiarcos et al., 2012)

[2] See (McCrae et al., 2012)

[3] `https://www.w3.org/2016/05/ontolex/`. See also for a kind of historical view on the development of *lemon* towards OntoLex-Lemon (McCrae et al., 2017).

[4] See for example (Declerck et al., 2017), (Khan et al., 2017) or (Tiberius and Declerck, 2017).

[5] See (Declerck and Lendvai, 2016).

[6] For example describing a lexicography module for OntoLex-Lemon. See (Bosque-Gil et al., 2017) and `https://www.w3.org/community/ontolex/wiki/Lexicography`.

[7] See again (McCrae et al., 2012).

[8] See again (McCrae et al., 2017).

to be applied to its core module, the ontology-lexicon interface, in order to potentially realise our goal. However, we will first discuss some observations made regarding the current status of the LLOD.

2. Observations on the current State of the Linguistic Linked Data Cloud

When looking at the current state of the Linguistic Linked Open Data (LLOD) in detail, which is displayed in Figure 1[9], it can be noticed that the data sets published in this cloud are classified along the lines of six categories:

- Corpora
- Terminologies, Thesauri and Knowledge Bases
- Lexicons and Dictionaries
- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases

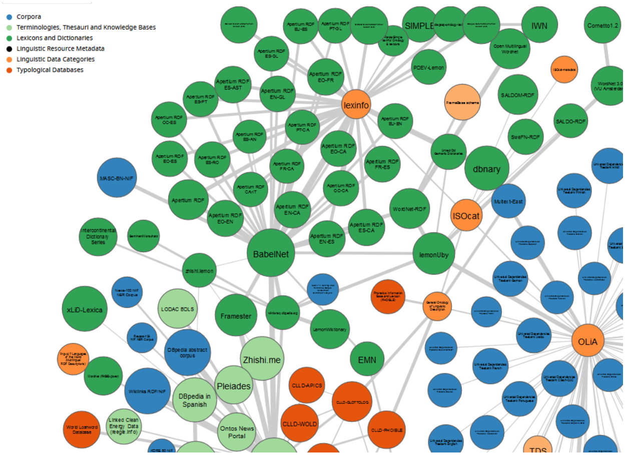

Figure 1: A partial view on the Linguistic Linked Open Data cloud, as of July 2017.

To access lexical items in the LLOD, it is easier thus to enter a lexicon or dictionary data set first and this probably reflects the meaning of the term (or ontology class) `LexicalEntry` that is used by the dictionaries or lexicons in the LLOD, which are making use of *lemon* or OntoLex-Lemon.

Here, we adopt the Wikipedia definition of "lexical entry", which states: "In lexicography, a lexical item (or lexical unit/ LU, lexical entry) is a single word, a part of a word, or a chain of words (=catena) that forms the basic elements of a language's lexicon (=vocabulary)."[10].

The question now is if the term (or ontology class) `ontolex:LexicalEntry` in OntoLex-Lemon only has a "lexicographic" acceptance (i.e an entry has to be part of a dictionary or a lexicon), and this applies even more if we consider that the modelling of language data that occur in the labels of a taxonomy or an ontology is done without taking into consideration any dictionary or lexicon.

We think that in this respect, the core module of OntoLex-Lemon should be clearly distinguished from the definition of a lexical entry that is to be provided by the upcoming lexicography module, which is currently being discussed within the W3C Ontology-Lexica community[11]. The participants in this discussion are perfectly aware of this issue, as they are suggesting the name "DictionaryEntry" to represent the structure of an entry in a (mostly non-LD-native) dictionary, and thus differentiating it from a "LexicalEntry", which is modelling a lexical item that is not necessarily included in a lexicographic work. This is in fact the view supported by OntoLex-Lemon, as the information about the naming of a collection (possibly a lexicon) of lexical items is left to the "LInguistic MEtadata" (*lime*) module, which describes metadata as related to the lexicon-ontology interface.[12].

Now turning our attention back to the analysis of the LLOD again, we consider the example of the aggregated RDF Apertium bi-lingual dictionaries[13] in greater detail. For the RDF version of Apertium, Spanish lexical data that were originally contained in different bi-lingual dictionaries have been merged into one data set and lexical entries of the source and the target languages are pointing to the same BabelNet synset[14]. BabelNet is developing a hub for references to senses and encyclopaedic sources in the LLOD. Having a source language word and a target language word pointing to the same BabelNet meaning (or sense) can therefore be considered a good way to indicate their appropriateness for a translation relation. This is a good case where we can see the benefit of the LLOD approach to the modelling and linking of language data.

At the same time, the Italian lexical data included in RDF Apertium, in its bi-lingual Catalan-Italian dictionary, does not have any link to the Italian data included in the SIMPLE lexicon[15]. Furthermore, SIMPLE is not linking to BabelNet, but to another source containing senses. The direct question is here: why do we have two "entries" for one and the same (Italian) word, in SIMPLE and in Apertium[16]?

When looking at the corresponding RDF Apertium and

[9] The full LLOD cloud can be accessed at `http://linguistic-lod.org/llod-cloud`. There, one can click on the various nodes and get more details about the data sets represented by the "bubbles".

[10] See `https://en.wikipedia.org/wiki/Lexical_item`.

[11] See (Bosque-Gil et al., 2017) and the on-line discussion at `https://www.w3.org/community/ontolex/wiki/Lexicography` for more details.

[12] See (Fiorelli et al., 2015) and `https://www.w3.org/2016/05/ontolex/\#metadata-lime` for more details.

[13] Apertium is an open-source machine translation platform (see `https://www.apertium.org/index.eng.html`).For the porting of Apertium resources to *lemon* and their publication on the LLOD, see `https://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/`.

[14] See `http://babelnet.org/`.

[15] See `http://catalog.elra.info/product_info.php?products_id=881` for the original SIMPLE lexicon.

[16] In Apertium, information related to the word "bocca" (mouth)

SIMPLE data in the LLOD, the reader can observe that there are enough similar elements in each representation of the same lexical element. The main difference resides in the (way of) linking to a source representing the corresponding sense(s). One can ask then if, similar to the successful merging exercise done in the case of the monolingual Spanish lexicon in RDF Apertium, it would not be possible to merge all the triples into the LLOD dealing with the Italian form "bocca". In doing so, the merging would not lead to a specific lexicon, but to a data set itself, containing or linking to all related data or information/knowledge. This is basically what we would understand by a Linked Lexical Data cloud.
In the following section, we propose a short analysis of the current version of OntoLex-Lemon.

3. Observations on the current State of OntoLex-Lemon

The graphical view presented in Figure 2 demonstrates the organisation of the core module of OntoLex-Lemon: the "ontology-lexicon interface" (ontolex).

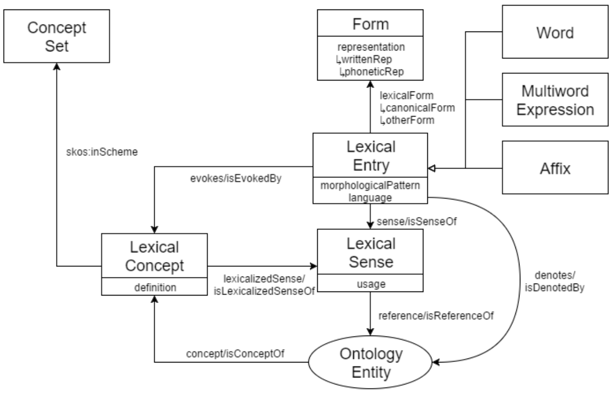

Figure 2: The core module of OntoLex-Lemon: Ontology Lexicon Interface. Graphic taken from `https://www.w3.org/2016/05/ontolex/`.

Looking now at the `LexicalEntry` class, it can be noticed that this class has kind of a pre-eminent position, which is not due to its central position in the graphic. The pre-eminence I see is the fact that none of the other elements in the field of morphosyntax information has a relation to sense, conceptual or referential resources. Therefore, they all have to "communicate" first with the class `LexicalEntry`. But as quoted before from (Gracia et al., 2017), we would prefer to see all elements of the model being first-class citizens. As a consequence of that, the resulting question is why an instance of a `ontolex:Form`, for example, cannot have a property linking to a sense or to an ontological reference.
One example which has recently been discussed[17] was the Spanish word "cura", which in English can mean "cure", when used in feminine, or "priest" (or similar), when used in masculine. One option for this would be to introduce two separate lexical entries with their corresponding canonical form and sense(s). Like this, the introduction of an instance of `LexicalEntry` would not only be motivated by the part of speech of the word to be represented, but also by its gender. And in addition to that, the sense would play a role in the decision on adding an entry or more for one word. I see in this a weakening principle of the modularity principle existing between the fields of lexical entries and lexical senses. An alternative solution would be to have only one "entry" for the Spanish noun "cura" and to allow the different canonical forms (one in feminine, one in masculine) to have a direct link to the corresponding sense(s). This way, we do not duplicate the number of entries, while keeping the same number of forms, and the OntoLex-Lemon elements (or classes) `LexicalEntry` and `Form` are being treated equally.
We extent this question to elements of the "Decomposition" module[18], which is displayed in Figure 3. This module supports the representation of components of a decomposed compound word or the components of a multi words expression.

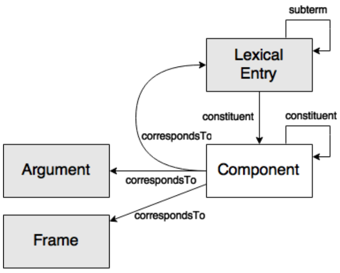

Figure 3: The Decomposition Module of OntoLex-Lemon. Graphic taken from `https://www.w3.org/2016/05/ontolex/`.

The cases we were dealing with are the German words "Erdöl" (*oil*) or "Erdgas" (*gas*) on the one hand and "Erdbeer" (*strawberry*) on the other. After decomposition, we have the components "Erd", which can be linked via the property `correspondsTo` to the OntoLex-Lemon entry "Erde" (*earth*), but it can be observed that "Erd" on its own is not a correct word or form of German. In one case, we now need to link the meaning of the component "Erd" to the sense representing a geological surface that needs to be drilled in order to extract oil (or gas), and in the other case to an agricultural sense of "Erd". We do not see how to do this if one has to link to the corresponding "Erde" entry first and we do not want to augment the number of "Erde" entries for this. Another option would be to add a restriction pointing to the corresponding component in the sense description, but then, we would have a direct link between a sense and a component (which is not a lexical entry or even a lexical form).

is available at `http://linguistic.linkeddata.es/page/id/apertium/lexiconIT/bocca-n-it` and in SIMPLE at `http://www.languagelibrary.eu/owl/simple/psc/pscLemon.ttl`.

[17] `https://www.w3.org/community/ontolex/wiki/Lexicography`.

[18] `https://www.w3.org/2016/05/ontolex/\#decomposition-decomp`.

There seem to be enough cases that call for a loosening of the current restriction allowing that only a LexicalEntry can be linked by a property to a LexicalSense, a LexicalConcept or even an ontological reference.

In doing so, the model would be very close to the already quoted statement that "every lexical element (headword, sense, written form, grammatical at- tribute, etc.) is treated as a first-class citizen, being identified by its own URI at a Web scale, and being attached to its own descriptive information and linked to other relevant elements through RDF statements" (Gracia et al., 2017).

4. About the Status of LexicalEntry in OntoLex-Lemon

The discussion about the problematic cases resulting from the fact that the class LexicalEntry is playing a central (or pivotal) role as an intermediate between morpho-syntactic and semantic descriptions of lexical data leads to the fundamental question about its status within the model. Looking at many examples of encoding of entries with *lemon* or OntoLex-Lemon, one gets the impression that an instance of the LexicalEntry class is in fact a grouping of related word forms, based on their shared Part-of-Speech information. Is this the case, the labelling of the class in term of LexicalEntry would be misleading. I am wondering if in such a graph-based model, in which all nodes are to be considered "first-class citizens" (Gracia et al., 2017), such a class as LexicalEntry is still needed. In fact, the labelling of this class seems to be a reminiscence of non LD-native dictionaries, in which the access to lexical data was guided by lexical entries, that were organized by extra-linguistic principles, as this is for example the case for the alphabetic ordering of entries, which is "an arbitrary system which brings together completely unrelated words in sequences like: *redneck*, *redness*, *redo*, *redolent*, *redoubtable*" (Rundell, 2015).

5. Towards a Linked Lexical Data cloud

As certain professional lexicographers are aiming at an e-lexicography beyond dictionaries[19], is it not appropriate to also consider an e-lexicography beyond lexical entries, but dealing only with lexical data that can be directly linked to each other in a huge network, which we would like to call the Linked Lexical Data cloud. In this cloud the different lexical data could be linked not only to each other but also to other types of data, and be directly integrated in LLOD-based applications. One could also aim at merging lexical data and so to reduce redundancies of data descriptions.

In this Linked Data Lexical cloud, both the users and Natural Language Processing applications would have direct access to the needed lexical information, responding thus to a certain extend to the needs formulated by publishers and other professionals in the e-lexicographic field. (Køhler Simonsen, 2017) for example stresses the fact that "[...] the biggest problem of lexicography is that lexicographic products are no longer perceived as relevant for the vast majority of people. Most people in fact do not use dictionaries, and if they need to find help when communicating or when looking for data, they simply use the Internet instead. So dictionaries are in fact not being used as much as we want them to be. The most important question is: why do not people use online or mobile dictionaries? Obviously, there are a number of reasons, but I would argue that the most important reason is that most lexicographic resources are not tool-integrated and not specifically related to the user's job tasks". In order to be able to implement business models for the modern e-lexicography, (Køhler Simonsen, 2017) requires among others that lexicographic products are moving to lexicographic services, the integration of lexicographic data in lexicographic platform and distribution, and to take increasingly into account the lexicographic users and their needs, and in general a move "from dictionary to lexicographic data in software [and] artificial intelligence". The intended Linked Lexical Data cloud could be instrumental in reaching those goals.

6. Conclusion

We presented some considerations about the current state of the Linguistic Linked Data (LLOD) cloud and the OntoLex-Lemon model, which is a core component of the LLOD. As we argue that within the LLOD it would be beneficial to have a formal representation and a dense linking of lexical data that are not necessarily included in a lexicon or in a dictionary-based data-set, we end up in suggesting slight modifications of the OntoLex-Lemon model, also on the base of the discussion of some examples that are difficult, if not impossible, to model with the current version of OntoLex-Lemon. While the suggested modifications of OntoLex-Lemon are minimal, they lead to a fundamental question on the status of the texttttLexicalEntry class, which ultimately could be made optional or disappear, at least in the context of the intended Linked Lexical Data cloud. We presented also briefly some views proposed by professionals in the field of lexicography publishing, and which are in line with our consideration that dictionaries are no longer needed as container of lexical data in a Linked Data-based framework.

Acknowledgement

The DFKI contribution to this work has been partially funded by the German Federal Ministry of Education and Research under the funding code 01-W17001 for the project "DeepLee - Tiefes Lernen für End-to-End-Anwendungen in der Sprachtechnologie". Responsibility for the content of this publication is with the author. The ACDH contribution to this work has been partially funded by the H2020 project "ELEXIS" with Grant Agreement number 731015. We would like to thank the anonymous reviewers for their very helpful comments on the original submission of this paper. we also thank Eileen Schnur for her proof-reading on the first three sections of this paper.

[19] We borrow this expression from the title of a talk on "post-dictionary lexicography" given by Ilan Kernerman at eLex 2017, available at https://www.youtube.com/watch?v=yA3yg6wO5M8.

7. Bibliographical References

Bosque-Gil, J., Gracia, J., and Montiel-Ponsoda, E. (2017). Towards a module for lexicography in ontolex. In *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017), Galway, Ireland, June 18, 2017.*, pages 74–84.

Chiarcos, C., Hellmann, S., and Nordhoff, S., (2012). *Linking Linguistic Resources: Examples from the Open Linguistics Working Group*, pages 201–216. Springer Berlin Heidelberg, Berlin, Heidelberg.

Declerck, T. and Lendvai, P. (2016). Towards a formal representation of components of german compounds. In Micha Elsner et al., editors, *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Humboldt University, ACL, 8.

Declerck, T., Tiberius, C., and Wandl-Vogt, E. (2017). Encoding lexicographic data in lemon: Lessons learned. In John P. McCrae, et al., editors, *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets*. CEURS, 8.

Fiorelli, M., Stellato, A., McCrae, J., Cimiano, P., and Pazienza, M. T. (2015). LIME: the Metadata Module for OntoLex. In *Proceedings of 12th Extended Semantic Web Conference*.

Gracia, J., Kernerman, I., and Bosque-Gil, J. (2017). Toward linked data-native dictionaries. In Iztok Kosem, et al., editors, *Proceedings of the eLex 2017 conference*, pages 550–559. INT, TrojÃna and Lexical Computing, Lexical Computing CZ s.r.o., 9.

Khan, F., Bellandi, A., Boschetti, F., and Monachini, M. (2017). The challenges of converting legacy lexical resources to linked open data using ontolex-lemon: The case of the intermediate liddell-scott lexicon. In *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017), Galway, Ireland, June 18, 2017.*, pages 43–50.

Køhler Simonsen, H. (2017). Lexicography: What is the business model? In Iztok Kosem, et al., editors, *Electronic Lexicography in the 21st Century*, pages 395–415. Lexical Computing CZ s.r.o.

McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gomez-Perez, A., Garcia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4):701–719.

McCrae, J. P., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. In Iztok Kosem, et al., editors, *Proceedings of eLex 2017*, pages 587–597. INT, Trojína and Lexical Computing, Lexical Computing CZ s.r.o., 9.

Rundell, M. (2015). From print to digital: Implications for dictionary policy and lexicographic conventions. *Lexikos*, 25(1).

Tiberius, C. and Declerck, T. (2017). A lemon model for the anw dictionary. In Iztok Kosem, et al., editors, *Proceedings of the eLex 2017 conference*, pages 237–251. INT, Trojína and Lexical Computing, Lexical Computing CZ s.r.o., 9.