Interoperability of Language Data in lexicographic Frameworks and in Knowledge Systems

Thierry Declerck1, 2

1 DFKI GmbH, Multilingual Technologies

Stuhlsatzenhausweg, 3

66123 Saarbrücken, Germany

2 Austrian Centre for Digital Humanities; Austrian Academy of Sciences,

Sonnenfelsgasse, 19
1010 Vienna, Austria
E-mail: declerck@dfki.de

Abstract

In this short paper we present recent work aiming at achieving an interoperable formal representation of both knowledge and language data. Our presentation is grounded in language technology based initiatives that emerged in the context of the expanding Linked Open Data framework and which led to the creation of the so-called Linguistic Linked Open Data cloud. We see in this development a possibility for establishing a new bridge between lexicographic resources and (domain specific) knowledge data sets.

Keywords: Linguistic Linked Open Data; Knowledge Bases; eLexicography

1. Introduction

The discussion on the relation between the knowledge of the *words* and the knowledge of the *world* is for sure not new and is at the core of the decision on how to encode entries in dictionaries and encyclopaedias. While we cannot here list all positions on this issue and all definitions of the meaning(s) of words, we would like to mark our agreement with a statement made by Hobbs J.R. (1987: 6): “An old favorite question for lexical theorists is whether one can make a useful distinction between linguistic knowledge and world knowledge. The position I have articulated leads one to an answer that can be stated briefly. There is no useful distinction.” Our sympathy for this statement lies in the fact that in the overwhelming number of cases knowledge is transmitted by natural language, so that a good understanding of such expressions is a necessary basis for acquiring and using knowledge. Now, we need to make sure that this postulated similarity of the two types of knowledge is made operable. Therefore we are focusing on recent work dealing with the common representation of knowledge and language data. We restrict ourselves to the consideration of knowledge as structured data encoded in the context of the expanding Linked Open Data cloud[[1]](#footnote-1), which is based on the use of W3C standards, a set of formal languages developed originally for the Semantic Web initiative[[2]](#footnote-2). An approach for marking explicitly the relation between language data and knowledge objects consists in analysing and formally representing the language data used in such knowledge objects. The model for this formal representation of language data is briefly described in the next section.

1. Lexicon Model for Ontologies

In recent years we have experienced a trend towards the formal representation of language data in order for those to be integrated in the Linked (Open) Data cloud (LOD). As a result of this, even a Linguistic Linked Open Data (LLOD) cloud[[3]](#footnote-3) has been created, and is constantly increasing. At the basis of this LLOD the use of various representation formats has been tested and implemented, being SKOS, SKOS-XL[[4]](#footnote-4) or the lexicon model for ontologies (*lemon*), resulting from the W3C Ontology-Lexica Community Group[[5]](#footnote-5). The latest has the big advantage of being designed for the encoding of all possible lexical and grammatical features; the only requirement being that those features are represented with the help of the W3C standards in use on the LOD framework. Applying *lemon* to all language data associated to knowledge objects, we can obtain a precise mapping not only of the words used, but also of their lexical and grammatical features to objects encoded in the knowledge data sets in the LOD.

In a sense, we are performing something similar to the so-called distributional semantics approach[[6]](#footnote-6). But while in this approach the contexts of the words to be semantically represented, in the form of vectors, are sequences in large corpora, the contexts we are considering for the semantic encoding of language data are given by elements of structured knowledge objects (thesauri, taxonomies, ontologies). Here we can paraphrase Firth[[7]](#footnote-7); whose famous sentence “You shall know a word by the company it keeps” can be applied to our case, in which the “company” is in fact a large set of knowledge data sets.

A major difference between the distributional approach and the one we describe here is the fact that we take into account a fine-grained description of relevant lexical features of words, giving to the lexicographer the possibility to select more easily the words that should be considered for integration in dictionaries. And the pointing to structured knowledge objects for marking the semantics of the words is probably more intuitive for user of language resources as the high-dimensional vectors resulting from the distributional approach[[8]](#footnote-8). In this we are following the recommendations for linguistically grounded ontologies, as those are formulated in (Buitelaar et al., 2009).

Below we display the graphical representation of the core model of *lemon*, which is also called “OntoLex” (Ontology-lexicon interface).



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

The core model is accompanied by four other modules:

* Syntax and Semantics (synsem)
* Decomposition (decomp)
* Variation and Translation (vartrans)
* Linguistic Metadata (lime).

In the core module (ontolex), the reader can see how lexical features can be linked by appropriated relation markers (properties) to knowledge objects outside of the description of lexical features properly speaking. Those objects can be contained in taxonomies or ontologies on the one hand, or are part of a conceptual/mental representation of terms (in this case using then the SKOS framework).

The core model is simple and abstract. The detailed lexical features are not part of the *lemon* framework, but imported from the LexInfo set of descriptors[[9]](#footnote-9). And in fact this kind of import can be extended to data and features that are not yet used by *lemon,* and we expect the lexicographic field to consider much more lexical features than those that are included in LexInfo.

1. Relevance for eLexicography

In this context, cooperation between the past LIDER project[[10]](#footnote-10) and the ENeL Cost Action[[11]](#footnote-11) work has been established, for porting lexicographic data onto a LLOD compliant format. In various exercises it could be shown how lexicographic data could be linked not only to other language data but also to encyclopaedic knowledge data sets in the LOD[[12]](#footnote-12). In this, we explored ways for expanding the coverage of lexicographic work, allowing the lexicographers to work more focused on her/his field, while making information of related fields available in a machine readable way.

Now, at the same time, we saw that many knowledge data sets in the LOD are making extensive usage of a huge amount of (multilingual) language data. But not following any convention or encoding standards that would ease a principled access to these data for the purpose of lexicographic work. We need thus to discuss how this language data can be processed and transformed in such a way that the lexicographer can decide if and how to integrate word forms and knowledge associated to those language data in lexicographic reference works.

Last but not least: large number of (past and non-digital) dictionaries have been the carrier of much more information that just about the words, giving precious information on the cultural and historical dimension of the usage of certain words. It is certainly a very valuable challenge to extract this information from those very rich lexicographic works and to encode them in such a way that they can be accessed on and linked to the Linguistic Linked Data cloud.

1. Acknowledgements

The current results emerged from collaborative work by members of the Open Knowledge Foundation's Working Group on Open Data in Linguistics (http://linguistics.okfn.org/) and of the W3C Ontology-Lexica Community Group (https://www.w3.org/community/ontolex/), as well as from a fruitful cooperation established between the past LIDER project (http://www.lider-project.eu/), under number 610782, and the running ENeL COST Action (http://www.elexicography.eu/), under number IS1305.

1. References

Buitelaar, P., Cimiano, P., Haase, P. & Sintek, M. (2009). Towards Linguistically Grounded Ontologies. In: Proceedings of the European Semantic Web Conference (ESWC).

Chiarcos, C., Hellmann, S. & Nordhoff, S. (2012). Linking linguistic resources: Examples from the Open Linguistics Working Group, In: Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.), Linked Data in Linguistics. Representing Language Data and Metadata, Springer, Heidelberg, p. 201-216.

Declerck, T. & Wandl-Vogt, E. (2014). How to semantically relate dialectal Dictionaries in the Linked Data Framework. In: Kalliopi Zervanou, Cristina Vertan, Antal van den Bosch, Caroline Sporleder (eds.): Proceedings of the The 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2014), Gothenburg, Sweden, ACL.

Firth, J.R. (1957). "A synopsis of linguistic theory 1930-1955". Studies in Linguistic Analysis. Oxford: Philological Society: 1–32. Reprinted in F.R. Palmer, ed. (1968). Selected Papers of J.R. Firth 1952-1959. London: Longman

Hobbs, J-R. (1987). World Knowledge and Word Meaning. In Proceedings of the Third Workshop on Theoretical Issues in Natural Language Processing, TINLAP-3.

Sahlgren, M. (2008). "The Distributional Hypothesis". Rivista di Linguistica. 20 (1): pp. 33–53

Turney, P.D. Pantel, P. (2010) From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research, (2010), 37, 141-188.

1. https://en.wikipedia.org/wiki/Linked\_data [↑](#footnote-ref-1)
2. See https://www.w3.org/2001/sw/wiki/Main\_Page \_Page for more details [↑](#footnote-ref-2)
3. See http://linguistic-lod.org/llod-cloud and (Chiarcos et al., 2012) [↑](#footnote-ref-3)
4. See respectively https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html and https://www.w3.org/TR/skos-reference/skos-xl.html [↑](#footnote-ref-4)
5. https://www.w3.org/community/ontolex/ [↑](#footnote-ref-5)
6. See (Sahlgren, M., 2008). [↑](#footnote-ref-6)
7. See (Firth, J. R. 1957:11). [↑](#footnote-ref-7)
8. A very good description on the use of vectors for representing the syntax and the semantic of words in context is given in http://cs224d.stanford.edu/syllabus.html, which discusses the relevance of deep learning for natural language processing. [↑](#footnote-ref-8)
9. See http://lexinfo.net/ for downloading the full LexInfo ontology. [↑](#footnote-ref-9)
10. http://www.lider-project.eu/ [↑](#footnote-ref-10)
11. http://www.elexicography.eu/ [↑](#footnote-ref-11)
12. See for example (Declerck & Wnadl-Vogt, 2014) [↑](#footnote-ref-12)