Harmonizing Lexical Data for their Linking to Knowledge Objects in the Linked Data Framework

Thierry Declerck
DFKI GmbH,
Language Technology Lab
Stuhlsatzenhausweg, 3
D-66123 Saarbrücken,
Germany
declerck@dfki.de

Abstract

In this position paper we discuss some of the experiences we made in describing lexical data using representation formalisms that are compatible for the publication of such data in the Linked Data framework. While we see a huge potential in the emerging Linguistic Linked Open Data, also supporting the publication of less-resourced language data on the same platform as for mainstream languages, we are wondering if, parallel to the widening of linking language data to both other language data and encyclopaedic knowledge present in the Linked Data cloud, it would not be beneficial to give more focus more on harmonization and merging of RDF encoded lexical data, instead of establishing links between such resources in the Linked Data.

1 Introduction

In recent years a lot of initiatives have emerged towards the RDF based representation of language data and the hereby opened possibility to publish those data in the Linked Open Data (LOD) cloud[1]. This development has been leading to the establishment of a specialized Linked Data (LD) cloud for language data. The actual diagram of this rapidly growing Linguistic Linked Open Data (LLOD) framework[2] reflects the distinct types of language data that already exist in LOD compliant formats, supporting their publication in the cloud and enabling their cross-linking and their linking to other knowledge objects available in the LOD context.

And to further stress the importance of this development, the main conference in the field of language resources, LREC, has declared the LOD as one of the hot topics of its 2014 edition[3] and we can observe from the list of accepted papers and workshops/tutorials that indeed this is really a hot topic for the description of language resources.

Some projects and initiatives have been very active in this field, and we want to mention here only a few, like the LOD2 project[4], which released among others the NIF (**N**LP **I**nterchange **F**ormat)[5] specifications, or the Monnet project[6], which delivered the *lemon* model for the representation of lexical

[1] See http://linkeddata.org/
[2] See http://linguistics.okfn.org/resources/llod/
[3] http://lrec2014.lrec-conf.org/en/calls-for-papers/lrec-2014-hot-topics/
[4] See http://lod2.eu/Welcome.html
[5] See http://nlp2rdf.org/nif-1-0
[6] http://cordis.europa.eu/fp7/ict/language-technologies/project-monnet_en.html

data in ontologies[7], and the current project LIDER, which is aiming at providing "the basis for the creation of a Linguistic Linked Data cloud that can support content analytics tasks of unstructured multilingual cross-media content" [8]. Participants of those projects and many other researchers joined in standardization activities, mainly in the context of W3C, like the Ontolex community group[9].

We are also aware of works porting dialectal dictionaries (Wandl-Vogt and Declerck, 2013) or polarity lexicons (Buitelaar et al., 2013) onto LOD compliant representation formalisms. A benefit of such approaches is the fact that lexical data can be linked to meanings encoded in knowledge sources that are accessible via a URI, such as senses encoded in the DBpedia instantiation of Wiktioanry, and from there one can navigate to multilingual lexical equivalents, if those are available.

As a concrete example, working on historical German text, we could link the old word form "Fegfeur" (*purgatory*) via its modern German lemma "Fegefeuer" not only to a lexical sense in the DBpedia instantiation of Wiktionary: http://wiktionary.dbpedia.org/page/Fegefeuer-German-Noun-1de, also with access to 7 translations of this sense, but also leading to the DBpedia page for "purgatory", one get additional semantic information, so for example that the word is related to the categories "Christian_eschatology", "Christianity_and_death" etc.[10] And, in fact, the recent release of BabelNet 2.5 is summarizing this information in one page[11] for the reader, integrating information from WordNet, Wiktionary and Wikipedia. This example alone gives a very strong argument on why it is worth to encode language data using the same type of representation formalism as for knowledge objects available in the Linked Data cloud.

2 Representation Formalisms used

Based on the Resource Description Framework (RDF)[12], SKOS (Simple Knowledge Organization System)[13] "provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary."[14] This representation language is being widely used, since SKOS concepts can be (1) "semantically related to each other in informal hierarchies and association networks", (2) "the SKOS vocabulary itself can be extended to suit the needs of particular communities of practice" and finally, because it (3) "can also be seen as a bridging technology, providing the missing link between the rigorous logical formalism of ontology languages such as OWL and the chaotic, informal and weakly-structured world of Web-based collaboration tools."[15] With the use of SKOS (and RDF), we are also in the position to make language resources compatible with other language resource available in the LOD cloud, as we could see with our examples above with the DBpedia instantiation of Wiktionary16 or the very recent release of BabelNet. Since, contrary to most knowledge objects described in the LOD, we do not considers strings (encoding lemma and word forms as part of a language) as being just literals, but in also knowledge objects, we considered the use of SKOS-XL and of the lemon model, which was developed in the context of the Monnet project[17]. *lemon* is also available as an ontology[18].

3 A concrete Exercise with (German) polarity Lexicons

Inspired by (Buitelaar et al., 2013) we aimed at porting German polarity lexicons to a Linked Data compliant format, and so publish our data directly in the cloud. Our starting points are the following resources:

---

[7] See http://lemon-model.net/
[8] See http://www.lider-project.eu/
[9] http://www.w3.org/community/ontolex/wiki/Main_Page
[10] Details of this work is decribed in (Resch et al., 2014)
[11] See http://babelnet.org/search.jsp?word=Fegefeuer&lang=DE
[12] http://www.w3.org/RDF/
[13] http://www.w3.org/2004/02/skos/
[14] http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/
[15] Ibid.
[16] See http://dbpedia.org/Wiktionary. There, *lemon* is also used for the description of certain lexical properties.
[17] See http://lemon-model.net/
[18] See http://www.monnet-project.eu/lemon

- A polarity lexicon for German[19] (Clematide and Klenner, 2010)
- GermanPolarityClues[20] (Waltinger, 2010)
- GermanSentiSpin[21]
- SentiWS[22] (Remus et al., 2010)

3.1 Pre-Processing of the lexical Data: Harmonization

As the reader can imagine, all those resources were available in distinct formats and containing distinct types of features. Therefore, we first had first to define a pre-processing of the different lexical data for the purpose of their harmonisation. This point leads us to a general remark: It is by far not enough to transform the representation of the lexical data onto RDF and related languages for ensuring their semantic interoperability in the LOD cloud, but preliminary work has to be performed. Just to give an example of the outcome of this work, we present a harmonized entry in Figure 1 below:

```
"fehler" => {                                                   # lemma
        "prov::GermanPC.lex" => {                               # provenance info
                "pos::N" => {                                   # PoS info
                        "pol_rank" => "0.783019",               # ranking in the orginal source
                        "pol_val" => "NEG",                     # polarity feature in the orig souce
                },
        },
        "prov::GermanSentiSpin.lex" => {
                "pos::N" => {
                        "pol_rank" => "0.0087112",
                        "pol_val" => "NEG",
                },
        },
        "prov::GermanSentiWS.lex" => {
                "pos::N" => {
                        "pol_rank" => "0.6752",
                        "pol_val" => "NEG",
                },
        },
        "prov::german.lex" => {
                "pos::N" => {
                        "pol_rank" => "0.7",
                        "pol_val" => "NEG",
                },
        },
},
```

Figure 1: The harmonized entry “fehler” (*error*) . The remaining differences in this polarity lexicon can be only in the value of the features “pos”, “pol_val” and “pol_rank”.

Only on the base of this harmonized lexicon, we started to model the lexical resource for publication on the LOD framework. But before getting onto the presentation of the model, we should note that the harmonized lexicon also contributed to a reduction of the lexical data: instead of originally 4 (lemma) entries, we have now only one.

[19] Downloadable at http://sentimental.li/german.lex
[20] Downloadable at http://www.ulliwaltinger.de/sentiment/
[21] SentiSpin is originally an English resource (Takamura eta., 2005), tranlsated to German by (Waltinger, 2010b).
[22] Downloadable at http://asv.informatik.uni-leipzig.de/download/sentiws.html

3.2 The LOD compliant Representation of the harmonized polarity Lexicon

Our work consisted in providing a representation of the lexical data using as much as possible information that is available in external resources, like the ISOcat registry[23], and an ontological model for the representation of polarity data, which is a slight extension of the MARL model, described in (Westerski et al., 2013). In below, we just display an excerpt of the description of the entry "fehler":

```
:LexicalSense_Fehler
      rdf:type lemon:LexicalSense ;
      rdfs:label "fehler"@de ;
      lemon:reference <http://wiktionary.dbpedia.org/page/Fehler-German-Noun-1de> .

:Opinion_Fehler
      rdf:type skosxl:Label , :lemma ;
      rdfs:label "Fehler"@de ;
      hasOpinionObject :Opinion_Fehler_2 , :Opinion_Fehler_3 , :Opinion_Fehler_4 ,
:Opinion_Fehler_1 ;
      :hasPoS <http://www.isocat.org/rest/dc/1333> ;
      skosxl:literalForm "fehler"@de .

:Opinion_Fehler_1
      rdf:type :Opinion_Object ;
      rdfs:label "Fehler"^^xsd:string ;
      op:assessedBy <http://tutorial-topbraid.com/lex_tm#german.lex> ;
      op:hasPolarity op:Negative ;
      op:maxPolarityValue "1.0"^^xsd:double ;
      op:minPolarityValue "-1.0"^^xsd:double ;
      op:polarityValue "-0.7"^^xsd:double .


  ……
```

Figure 2: The RDF, SKOS-XL and lemon representation of the entry, with a link to an ontological framework representing polarity information. The various polarities given by the various sources are represented as "OpinionObjects".

As the reader can see, such representation can link the lexical information to a wide range of related information, and what in the context of former infrastructures for language resources was represented by a set of external metadata can be incorporated here directly in the choice of classes and properties. In fact, we do not need to encode the information that the entry has PoS Noun, since this information is encoded in the details of the reference in Wiktionary/DBpedia we are pointing to.

4 Some "philosophical" Comments

The work we described briefly in this position paper, as well as work performed by researchers for porting for example dialectal dictionaries onto the LOD compliant formats (see Wandl-Vogt & Declerck, 2013) show a real potential for publishing distinct types of lexical data in the cloud, and to make this data accessible for both humans and machine in a very principled way. As noted, the use of carefully selected (widely accepted) classes and properties in the representation of the lexical data can also replace the use of complex metadata sets: parts of those metadata sets being implicitly encoded in the semantic representation the lexical data.

This positive aspect should not hide the fact that, at least in our opinion, the community is not thinking enough in providing for harmonization of the original lexical data. In many cases the data sets in the Linguistic Linked Open Data are redundant, repeating for example many times the lemmas of lexical entries in the different types of data set. We think that similar to the ISOcat data category we could aim at having a "centralized" repository for lemmas of one language, so that this lemma is not repeated for example in Wiktionary, Lexvo[24] and many other data sets in the LLOD. We are wondering if, in

[23] See for example http://www.isocat.org/rest/dc/1333 for our selected ISOcat entry for the pos "noun".
[24] See http://www.lexvo.org/

the precise context of the LOD – linking lexical data to other data sets in the cloud – it would not be possible to have exactly one lexical data set for reach language. Figure **3** below sketches our intended model, taking as example terminology in the field of financial reporting.

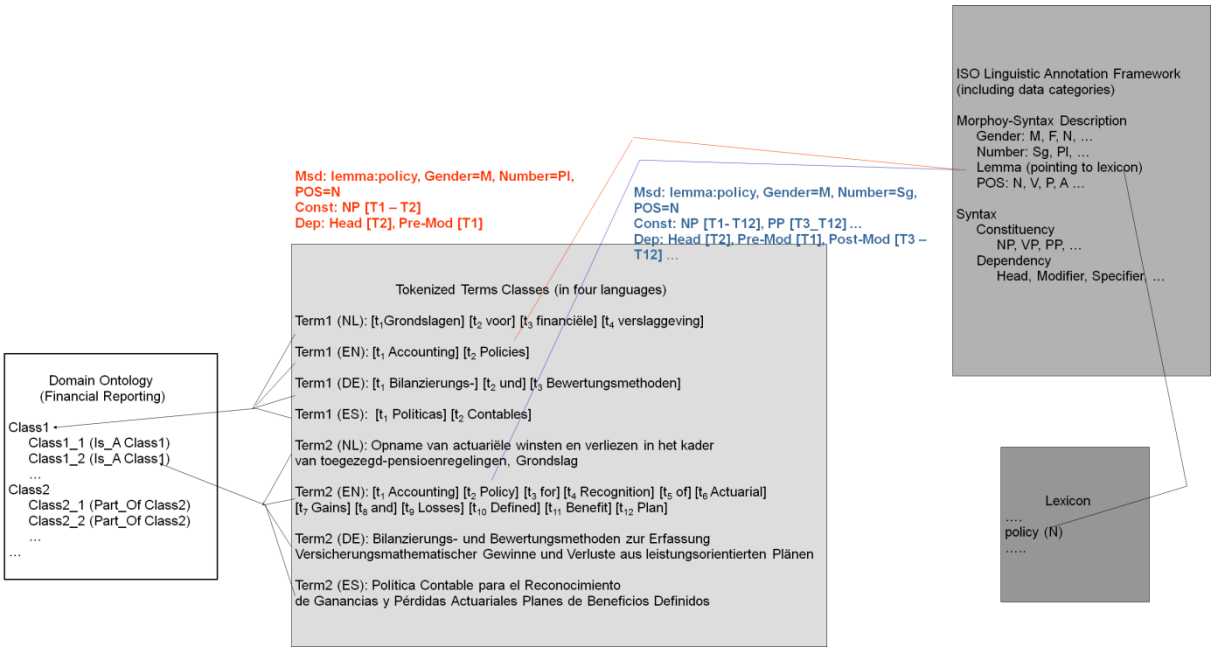

Figure 3: An example instantiation of the model we are aiming at: a unique (lemma) lexicon for one language (bottom right). Getting the full forms from a repository of such forms, including feature structures describing their morpho-syntactic information. Those are linking to occurrences of terms or labels that are used in knowledge objects (domain ontologies, taxonomies etc.). This model allows to precisely linking information from the lexicon, the morpho-syntactic descriptions and potential grammatical patterns as those are used in labels, comments or definitions in the context of knowledge objects in the LOD data sets. This model for representing lexical and linguistic data would be specialized for establishing linking between language data and representation of world knowledge. We expect from this approach an improvement in fields like domain specific machine translation and ontology-.based multilingual information extraction.

5 Conclusions

In this short position paper, we presented some experiences done in the context of the emerging Linguistic Linked Open Data framework. This lead us to make some comments on the way we could go for a much more "compressed" distribution of semantically (using LOD compliant representation languages) encoded language data, which could be more easily re-used in the context of knowledge-based NLP applications. The result would be a set of language specific "centralised" repositories of lemmas and related full forms, all equipped with URIs, that are used in the context of knowledge objects present in the Linked Data framework.

Acknowledgments

The research described in this paper is partly supported by the European Lider project. LIDER: "Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe" is a FP7 project with reference number 610782.

References

Buitelaar, P., Mihael Arcan, Carlos A. Iglesias, J. Fernando Sánchez, Carlo Strapparava (2013) Linguistic Linked Data for Sentiment Analysis. In: 2nd Workshop on Linked Data in Linguistics (LDL 2013): Representing and linking lexicons, terminologies and other language data. Collocated with the Conference on Generative Approaches to the Lexicon, Pisa, Italy

Clematide, S, Klenner, M. (2010). "Evaluation and extension of a polarity lexicon for German". In: Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA); Held in conjunction to ECAI 2010 Portugal, Lisbon, Portugal, 17 August 2010 - 17 August 2010, 7-13.

Clematide, S., Gindl, S., Klenner, M., Petrakis, S., Remus, R., Ruppenhofer, J., Waltinger, U. and Wiegand, M. (2012). MLSA ? A Multi-layered Reference Corpus for German Sentiment Analysis." In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, 23 May 2012 - 25 May 2012.

Hellmann, S., Lehmann, J., Auer, A. and Brümmer, M.**:** *Integrating NLP using Linked Data* In: 12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia.

Klenner, M., Clematide, S., Petrakis, S., Luder, M. (2012). "Compositional syntax-based phrase-level polarity annotation for German". In: The 10th International Workshop on Treebanks and Linguistic Theories (TLT 2012), Heidelberg, 06 January 2012 - 07 January 2012,

McCrae,J., Aguado-de-Cea, G., P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, T. Wunner.(2012) Interchanging lexical resources on the Semantic Web.*Language Resources and Evaluation*.

Remus, R., Quasthoff, U. and Heyer, G. (2010). "SentiWS - a Publicly Available German-language Resource for Sentiment Analysis." In: Proceedings of the 7th International Language Ressources and Evaluation (LREC'10), 2010

Resch, C., Declerck, T., Mörth, K, and Czeitschner, U. (2014) Linguistic and Semantic Annotation in Religious Memento Mori Literature. In *Proceedings of the 2nd Workshop on Language Ressources and Evaluation for Religious Texts.*

Hiroya Takamura, Takashi Inui, and Manabu Okumura. (2005). Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*.

Waltinger, U. (2010b). "Sentiment Analysis Reloaded: A Comparative Study On Sentiment Polarity Identification Combining Machine Learning And Subjectivity Features". In: Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST '10).

Wandl-Vogt, E., Declerck, T**.** (2013). Mapping a Traditional Dialectal Dictionary with Linked Open Data**.** In. Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) 2013. *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut.

Westerski, Adam and Sánchez-Rada, J. Fernando, Marl Ontology Specification, V1.0 May 2013, available at http://www.gsi.dit.upm.es/ontologies/marl