

# Standardized Multilingual Language Resources for the Web of Data: <http://corpora.uni-leipzig.de/rdf>

Matthias Quasthoff<sup>1</sup>, Sebastian Hellmann<sup>2</sup>, Konrad Höffner<sup>2</sup>

<sup>1</sup>) Hasso Plattner Institute, Potsdam, [firstname.lastname@hpi.uni-potsdam.de](mailto:firstname.lastname@hpi.uni-potsdam.de),

<sup>2</sup>) University of Leipzig, [lastname@informatik.uni-leipzig.de](mailto:lastname@informatik.uni-leipzig.de)

## 1 Introduction

Statistical knowledge on natural languages is inevitable for various kinds of services requiring Natural Language Processing (NLP) functionality, such as information retrieval. The NLP Group at the University of Leipzig started providing such statistical information for more than 50 languages in the Leipzig Corpora Collection (LCC) [1] more than a decade ago. Some of their corpora contain more than 5 million words and more than 300 million links between them, resulting in an accumulated size of about 60 million words and 814 million links in all corpora. So far, these valuable information could be accessed in a human-readable Web site and through a SOAP Web service, and excerpts of the data could be downloaded as SQL data dumps. A linked data interface for the LCC has now become desirable in order to allow a wider range of applications to make use of the corpora.

In this report, the LCC linked data interface is presented. This new service provides information about almost 60 million resources in approximately 900 million triples. Additionally, links to other vocabulary such as WordNet [2] and to DBpedia [3] are offered. The service is realized using a customized version of D2R Server [4].

## 2 Triplification of corpus data

### 2.1 Corpus vocabulary

For each language available, a separate *lcc:Corpus*<sup>1</sup> dataset is described, among others using the VoiD<sup>2</sup> vocabulary. Each corpus contains descriptions of *lcc:Words*, such as information on sentence- and neighbor-based *lcc:cooccurrences*, *lcc:frequency* etc. For the German corpus, additional information about semantic relations between words are available, e.g., whether two words are synonyms, antonyms, meronyms etc. Due to the different goals of LCC and WordNet, WordNet's

<sup>1</sup> <http://corpora.uni-leipzig.de/rdf/schema/0.1>

<sup>2</sup> <http://rdfs.org/ns/void>

semantic relations (such as *wn:antonymOf* or *wn:meronymOf*) cannot be used on *lcc:Words* directly. Instead, corresponding LCC relations are linked to the WordNet vocabulary using the formal rule “if two *lcc:Words* belong to a specific semantic LCC relation, each of the words belongs to anonymous *wn:Synsets*, and these synsets belong to the corresponding WordNet relation.”

We created an initial mapping from LCC to DBpedia. For this mapping we followed the general paradigm of creating few links with a high accuracy, based on matching labels of *lcc:Words* and DBpedia entries. All in all we created 117.629 links for the English corpus, 58.618 of which link to DBpedia disambiguation pages. A more versatile *m* to *n* mapping resolving polysemy, synonymy etc. is subject to further research. The existing mapping already proves valuable as it adds some semantic meaning to LCC words, as these so far are basically strings with statistical meta-data, and DBpedia is considered a hub in the Web of Data linking to other knowledge stores.

## 2.2 Triplification software

Because of the size and structure of the LCC corpora, D2R Server needed to be modified to efficiently publish them as linked data. First, the off-the-shelf version would have generated more than one million statements for high-frequency words. Second, due to D2RQ’s formal relational algebra, a prohibitive amount of join queries would have been issued against the LCC database. Thus, the D2RQ model implementation has been replaced with a composition of object-relational and object-triple mapping. The resulting RDF-OO-RDF translation is much more efficient than the comparable direct translation provided by D2RQ, at the cost of having to configure two mappings (from relational database schema to domain object model, and from object model to RDF.) The source code of the solution can be downloaded from the service’s Web site<sup>3</sup>.

## 3 Conclusion

In this article, we presented a straight-forward approach on how to publish standardized, multi-lingual language resources as linked data, and implemented the approach for more than 40 languages from all continents. Using this tool set, any interested party can publish their own corpora as linked data, e.g., for other languages, or different scientific, legal, or cultural domains. With our approach of using multi-lingual RDF labels of DBpedia entries, some common understanding is added to the corpora, which by design are agnostic of any potential semantics.

<sup>3</sup> <http://corpora.uni-leipzig.de/rdf/d2r-server-0.6-corpora.zip>

## References

1. Richter, M., Quasthoff, U., Hallsteinsdottir, E., Biemann, C.: Exploiting the leipzig corpora collection. In: Proc. of the IS-LTC 2006. (2006)
2. van Assem, M., Gangemi, A., Schreiber, G.: Rdf/owl representation of wordnet. <http://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/> (2006)
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007). Springer (2008) 722–735
4. Bizer, C., Seaborne, A.: D2rq – treating non-rdf databases as virtual rdf graphs. In: Proc. of the 3rd International Semantic Web Conference, Springer (2004)