Representing Swedish Lexical Resources in RDF with *lemon*

Lars Borin[1] and Dana Dannélls[1] and Markus Forsberg[1] and John P. McCrae[2]

[1] Språkbanken, University of Gothenburg
{lars.borin, dana.dannells, markus.forsberg}@svenska.gu.se
[2] Cognitive Interaction Technology Center of Excellence, University of Bielefeld
jmccrae@cit-ec.uni-bielefeld.de

Abstract. The paper presents an ongoing project which aims to publish Swedish lexical-semantic resources using Semantic Web and Linked Data technologies. In this article, we highlight the practical conversion methods and challenges of converting three of the Swedish language resources in RDF with *lemon*.

Keywords: Language Technology, Lexical Resources, Lexical Semantics.

1 Introduction

For state-of-the-art lexical resources, availability as Linked Open Data (LOD) is a basic requirement for their widest possible dissemination and use in research, education, development of products and services. In addition, in order to provide language support to individuals requiring augmentative and alternative communication (AAC) we need linguistic resources suitably organized and represented, e.g., sign language material, symbol and image libraries adapted to multiple cognitive levels, as well as textual support in many languages. So far, in Sweden, these resources have been developed as separate and uncoordinated efforts, either commercially or by non-profit organizations targeting specific groups and needs. In the long run, this is an exclusive and expensive way of proceeding, leading to limited usefulness. In the project, we aim to link Concept Coding Framework (CCF) technology and some symbol sets, to a common LOD format for languages resources (LRs) to be developed together with Språkbanken (The Swedish Language Bank),[3] which will be a great step forward.

There are ongoing international initiatives aiming to define suitable formats for publishing linguistic content according to linked open data principles [1]. Integrating linguistic content on the Web by using these principles is central for many language technology applications. It requires harmonization on different levels in particular on the metadata level. In this paper we present our first attempt to publish three Swedish lexical-semantic resources in RDF with *lemon* [2].

[3] <http://spraakbanken.gu.se/>

2 *lemon*

lemon (Lexicon Model for Ontologies) is a model for associating linguistic information with ontologies,[4] in particular Semantic Web ontologies. The model builds on existing models for incorporating multilingual knowledge in ontologies, and for the representation of lexical resources [3]. *lemon* is built around the principal of *semantics by reference* [4]. It separates the lexical layer, that is the words and their morphology and syntactic behaviour, and the semantic layer in the ontology, which describes the domain-specific meaning of that entry. The model of *lemon* is based around lexical entries, which connect to ontology entities, by means of an object called *LexicalSense*, which refers to one meaning of a word, or correspondingly a pair consisting of the word and the meaning. In this sense, the model of *lemon* is primarily semasiological, i.e, organized around words, as opposed to the onomasiological resources, such as SALDO, which are primarily built around senses. However, the usage of the sense object and the distributed nature of the RDF graph model, means that from a linked data viewpoint this distinction is of less relevance, and *lemon* proves to be an effective model for the lexical resources discussed here.

The *lemon* model has since 2011 been the focus of the W3C OntoLex community group,[5] and as such significant developments on both the model and its applications are still active. In particular, *lemon* has already been used successfully for the representation of several existing lexical resources, most notably WordNet [5], UBY [6] and BabelNet [7]. Furthermore, the use of *lemon* has already proved to be a key component in systems for tasks such as question answering [8], natural language generation [9] and information extraction [10].

3 Converting the Swedish Lexical Resources into RDF with *lemon*

Språkbanken at the Department of Swedish, University of Gothenburg, Sweden maintains a very large collection of lexical resources for both modern and historical Swedish. Currently there exist 23 different lexical resources with over 700,000 lexical entries. Within the time frame of our project we so far considered three of the modern lexicons, which are also freely available in Lexical Markup Framework (LMF) [11]. As we will show in this chapter, the form of these lexical resources varies substantially. We minimize this variation with *lemon*. Since *lemon* is builds on LMF, it allows easy conversion supported by EXtensible Stylesheet Language XSL Transformation mechanism.[6]

SALDO, the new version of the Swedish Associative Thesaurus [12], is a semantically organized lexicon containing morphological and lexical-semantic information for more than 130,000 Swedish lexical entries of which 13,000 are verbs.[7] It is the largest freely available electronic Swedish lexical resource for language technology, and is the pivot of all the Swedish lexical resources maintained at Språkbanken. SALDO entries

[4] <http://lemon-model.net>
[5] <http://www.w3.org/community/ontolex/>
[6] <http://www.w3.org/Style/XSL/>
[7] <http://spraakbanken.gu.se/saldo>

are arranged in a hierarchical structure capturing semantic closeness between senses indicated by a unique sense identifier, in *lemon* this unique identifier is represented with the object *lemon:LexicalSense*. A challenge here was how to represent SALDO's *lemgram* which is a pairing of the word base form and its inflectional paradigm. *Lemgram* is represented with the object *lemon:LexicalEntry*. The base form is described with a lemma value of the lexical entry and is represented with the object *lemon:Form*. The inflectional paradigm is described with a form value combined with a digit, and is also represented with the object *lemon:Form*. We defined our objects for capturing the paradigm patterns and the morphosyntactic tags.

Swedish FrameNet (SweFN), created as a part of a large project called SweFN++ [13], is a lexical-semantic resource that has been expanded from and constructed in line with Berkeley FrameNet (BFN) [14].[8] It is defined in terms of semantic frames. Each frame is unique and is evoked by one or more target word(s) called *lexical unit* (LU) which carries valence information about the possible syntactic and semantic realizations of frame elements (FEs). Frames are represented with the object *lemon:LexicalSense*. The LUs evoked by a frame are linked to SALDO entities with the property *lemon:isSenseOf*. The property *lemon:SemArg* links to FEs objects. There are two types of FEs: *core* and *peripheral*, these are represented with the object *lemon:Argument* and are linked to either *uby:core* or *uby:peripheral* with the property *uby:coreType*. A challenge here was how to represent the syntactic and semantic realizations of FEs that are illustrated with annotated example sentences. In the LMF file they are annotated with extra, non-standardized tags. Our solution was to define our object *karpHash:example* to represent the annotated example sentences for each FE.

Lexin is a bilingual dictionary,[9] originally developed for immigrants by the Swedish national agency for education.[10] It contains detailed linguistic information for 15 languages including sentence and expression examples, sentence constructions, explanations through comments, synonyms, etc. A lexical entry in Lexin is represented with *lemon:LexicalSense*. Entries are linked to SALDO entries with *owl:sameAs* property. In addition, we defined the objects *spraakbanken:translation* and *spraakbanken:synonym* to represent translation equivalents of sentences in our RDF model.

4 Summary

We described the effort of transforming three Swedish lexical resources into LOD using Semantic Web and Linked Data technologies.[11] Deciding on how to transform the lexical resource attributes to *lemon* features has been carried out manually for each resource. Once the transformation is decided, the integration is conducted automatically. Publishing lexical resources in Swedish as RDF data is valuable for a variety of use cases. One of the many benefits of having this semantically interlinked content is to

[8] <http://spraakbanken.gu.se/eng/resource/swefn>
[9] <http://spraakbanken.gu.se/eng/resource/lexin>
[10] <http://www.skolverket.se/>
[11] The published resources can be accessed here: <http://spraakbanken.gu.se/rdf>

enhance accessibility and availability on the web, in particular for language technology applications.

Acknowledgements

The research presented here has been conducted with funding by VINNOVA (Swedish Governmental Agency for Innovation Systems; grant agreement 2013-04996), and by the University of Gothenburg through its support of the Centre for Language Technology.[12]

References

1. Chiarcos, C., Nordhoff, S., Hellmann, S., eds.: Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata. Springer (2012)
2. McCrae, J., Spohr, D., Cimiano, P.: Linking lexical resources and ontologies on the semantic web with lemon. In: The Semantic Web: Research and Applications. (2011) 245–259
3. Cimiano, P., Buitelaar, P., McCrae, J., Sintek, M.: Lexinfo: A declarative model for the lexicon-ontology interface. Web Semantics: Science, Services and Agents on the World Wide Web **9**(1) (2011)
4. Buitelaar, P. In: Ontology-based Semantic Lexicons: Mapping between Terms and Object Descriptions. Cambridge University Press (2010) 212–223
5. McCrae, J.P., Fellbaum, C., Cimiano, P.: Publishing and linking WordNet using RDF and lemon. In: Proceedings of the 3^{rd} Workshop on Linked Data in Linguistics. (2014)
6. Eckle-Kohler, J., McCrae, J., Chiarcos, C.: lemonUby-a large, interlinked, syntactically-rich resource for ontologies. Semantic Web Journal, submitted. (2014)
7. Ehrmann, M., Vannela, D., McCrae, J.P., Cecconi, F., Cimiano, P., Navigli, R.: Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC). (2014)
8. Unger, C., Cimiano, P.: Pythia: Compositional meaning construction for ontology-based question answering on the semantic web. In Munoz, R., ed.: Natural Language Processing and Information Systems: 16^{th} International Conference on Applications of Natural Language to Information Systems. Volume 6716., Springer (2011) 153–160
9. Cimiano, P., Lüker, J., Nagel, D., Unger, C.: Exploiting ontology lexica for generating natural language texts from RDF data. In: Proceedings of the 14^{th} European Workshop on Natural Language Generation. (2013) 10–19
10. Davis, B., Badra, F., Buitelaar, P., Wunner, T., Handschuh, S.: Squeezing lemon with GATE. In: Proceedings of the First Workshop on the Mulitlingual Semantic Web. (2011) 74
11. Francopoulo, G., George, M., Calzolari, N., Monachini, M., Bel, N., Pet, M., Soria, C., et al.: Lexical markup framework (LMF). In: International Conference on Language Resources and Evaluation LREC. (2006)
12. Borin, L., Forsberg, M., Lönngren, L.: SALDO: a touch of yin to WordNet's yang. Language Resources and Evaluation **47**(4) (2013) 1191–1211
13. Borin, L., Dannélls, D., Forsberg, M., Toporowska Gronostaj, M., Kokkinakis, D.: The past meets the present in Swedish FrameNet++. In: Proceedings of the 14^{th} EURALEX International Congress. (2010) 269–281
14. Fillmore, C.J., Johnson, C.R., Petruck, M.R.L.: Background to Framenet. International Journal of Lexicography **16**(3) (2003) 235–250

[12] <http://www.clt.se>