Towards a Linguistic Linked Open Data cloud: Linking the MASC

Christian Chiarcos

Information Science Institute, University of Southern California
chiarcos@daad-alumni.de

Abstract

I describe benefits of modeling linguistic resources as Linked Data, i.e., using RDF, publishing them under an open licence, and creating links between them. Further, an overview over currently on-going community efforts to create a Linked Open Data (sub-)cloud of linguistic resources will be given. Both aspects are illustrated for the MASC corpus.

1. Overview

Nowadays, computational linguistics, Natural Language Processing and Information Technology are confronted with an immense – and steadily growing – wealth of linguistic resources accumulated in more than half a century of computational linguistics (Dostert, 1955), of empirical, corpus-based study of language (Francis and Kučera, 1964), and of computational lexicography (Morris, 1969).

To make these resources available to the different communities interested in linguistic resources, and to facilitate their use, however, a number of technological challenges are to be addressed. One fundamental problem is the interoperability of existing language resources, a problem actively addressed by the community since the late 1980s (Text Encoding Initiative, 1990), but still a problem that is partially solved at best (Ide and Pustejovsky, 2010). A closely related challenge is information integration, i.e., how heterogeneous information from different sources can be retrieved and combined in an efficient way. To address both problems, the linguistic and NLP communities are developing generic standards for different types of linguistic resources, including the Lexical Markup Framework (Francopoulo et al., 2006, LMF) for lexical-semantic resources and the Graph Annotation Framework (Ide and Suderman, 2007, GrAF) for annotated corpora, both maintained by the ISO TC37/SC4.

Outside the linguistic community, similar problems have been addressed, for example, in the discussion of meta data for the world wide web. The formalisms proposed there eventually converged into the Resource Description Framework (Klyne et al., 2004, RDF, W3C recommendation 1999). RDF provides very generic data structures (labeled directed multi-graphs), that were applicable to a broader band-width of problems than originally anticipated. Hence, RDF was readily adopted in other domains, and employed for different tasks. Nowadays, RDF represents the state of the art of knowledge representation in many scientific disciplines, and eventually it became one of the fundamental elements of the Semantic Web. Because of its genericity, its further development was (and is) supported by a large and interdisciplinary community of developers and users, from academics as well as from industry. As a result, a rich technological ecosystem evolved, which includes different representation formats with varying degrees of compactness and readability (e.g., RDF/XML, RDF/Turtle, RDF/HDT),[1] specialized sub-languages for different tasks (e.g., RDFS for hierarchical structures, SKOS for semi-structured terminology bases, and OWL/DL for formally defined ontologies),[2] parsers, validators and (for OWL/DL) reasoners, several data bases (RDF triple stores) and query languages.

The potential of RDF for representing linguistic resources has long been recognized, in particular for lexical-semantic resources, where RDF can be employed to achieve interoperability between lexical resources with Semantic Web technologies (Gangemi et al., 2003), but also for linguistic corpora, where RDF technologies can be used to process, to store and to query multi-layer corpora (Burchardt et al., 2008).

In this talk, I briefly describe advantages of RDF for modeling linguistic resources, and in particular, linguistic corpora, using the Manually Annotated Sub-Corpus of American English (Ide et al., 2008, MASC) as an example. Aside from emphasizing the availability of infrastructures for efficiently storing and querying RDF data, I focus on two aspects, **interoperability** between different types of language resources, and **integration of information**. RDF extends resource-type specific formalisms like GrAF or LMF in that it establishes interoperability and information integration not only *for* annotated corpora or lexical-semantic resources, but also *between* both types of resources. Certainly, the LMF and the GrAF data model will guide the future development of standards for linguistics, but adding RDF as another possible serialization of these data models (along with classical XML linearizations) may open up the possibility to benefit from RDF infrastructures for specific tasks such as data integration, storing and querying. For the future of GrAF, this may mean that it evolves in a similar way as the LMF, i.e., that it gains a status as meta model for which multiple, but convertible linearizations in different formats are provided.

Interoperability of RDF data and information integration involve the **Linked (Open) Data paradigm** (Berners-Lee, 2006) that postulates four rules for the publication and representation of web resources: (1) Referred entities should be designated by using URIs, (2) these URIs should be resolvable over HTTP, (3) data should be represented by

[1] `http://www.w3.org/TR/rdf-syntax-grammar`, `http://www.w3.org/TR/turtle`, `http://www.w3.org/Submission/HDT`

[2] `http://www.w3.org/TR/rdf-schema`, `http://www.w3.org/TR/skos-reference`, `http://www.w3.org/TR/owl2-overview`

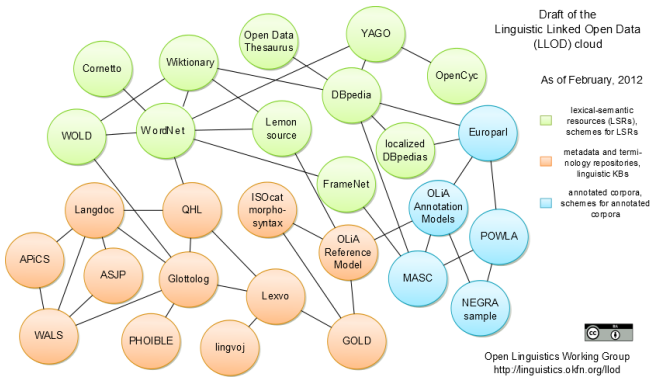

Figure 1: The Linguistic Linked Open Data (LLOD) diagram, draft version.

means of standards (such as RDF), (4) and a resource should include links to other resources. These rules establish information integration in that they require that entities can be addressed in a globally unambiguous way (1), that they can be accessed (2) and interpreted (3), and that entities that are associated on a conceptual level are also physically associated with each other (4).

The concept of Linked Data is closely coupled with the idea of **openness** (otherwise, the linking is only reproducible under certain conditions), and the definition of Linked Open Data has been extended with a 5 star rating system for data on the web. The first star is achieved by publishing data on the web (in any format) under an open license, the second, third and fourth star require machine-readable data, a non-proprietary format, and using standards like RDF, respectively. The fifth star is achieved by linking the data to other people's data to provide context.

If (linguistic) resources are published in accordance with these rules, it is possible to follow links between existing resources to find other, related data and exploit network effects. Following this insight, recent community efforts converge towards the development of a Linked Open Data (sub-)cloud of linguistic resources, the Linguistic Linked Open Data (LLOD) cloud, under the umbrella of the **Open Linguistics Working Group (OWLG)** of the Open Knowledge Foundation (Chiarcos et al., 2012). The OWLG is a multi-disciplinary network of researchers aiming to promote the idea of openness for linguistic resources, and dedicated to discussing and documenting the problems and benefits arising from open data in linguistics. It covers diverse disciplines, including language documentation, typology, computational linguistics, and information technology, just to name a few, and this diversity is also reflected in the current draft of the OWLG as illustrated in Fig. 1, which comprises general-purpose semantic knowledge bases (e.g., DBpedia), lexical resources (e.g., WordNet), annotated corpora (e.g., MASC), terminology repositories (e.g., an OWL linearization of the morphosyntactic profile of ISOcat), bibliographical data bases (e.g., Langdoc), and typological data bases (e.g., the World Atlas of Syntactic Structures, WALS).

I describe the integration of MASC in the LLOD cloud, and concrete use cases for the respective links.

2. References

Tim Berners-Lee. 2006. Design issues: Linked data. `http://www.w3.org/DesignIssues/LinkedData.html`. includes a 2010 addendum about linked *open* data.

Aljoscha Burchardt, Sebastian Padó, Dennis Spohr, et al. 2008. Formalising multi-layer corpora in OWL/DL – Lexicon modelling, querying and consistency control. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*, Hyderabad, India, Jan.

Christian Chiarcos, Sebastian Hellmann, Sebastian Nordhoff, et al. 2012. The Open Linguistics Working Group. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, May.

Leon E. Dostert. 1955. The Georgetown-IBM experiment. In William N. Locke and Andrew D. Booth, editors, *Machine Translation of Languages*, pages 124–135. John Wiley & Sons, New York.

W. Nelson Francis and Henry Kučera. 1964. Brown Corpus manual. Manual of information to accompany A standard corpus of present-day edited American English, for use with digital computers. Technical report, Brown University, Providence, Rhode Island.

Gil Francopoulo, Monte George, Nicoletta Calzolari, et al. 2006. Lexical Markup Framework (LMF). In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 233–236, Genoa, Italy.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. 2003. Sweetening WordNet with DOLCE. *AI magazine*, 24(3):13.

Nancy Ide and James Pustejovsky. 2010. What does interoperability mean, anyway? Toward an operational definition of interoperability. In *Proceedings of the 2nd International Conference on Global Interoperability for Language Resources (ICGL 2010)*, Hong Kong, China.

Nancy Ide and Keith Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proceedings of the 1st Linguistic Annotation Workshop (LAW 2007)*, pages 1–8, Prague, Czech Republic.

Nancy Ide, Collin F. Baker, Christiane Fellbaum, et al. 2008. MASC: The Manually Annotated Sub-Corpus of American English. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.

Graham Klyne, Jeremy J. Carroll, and Brian McBride. 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax. Technical report, W3C Recommendation.

William Morris, editor. 1969. *The American Heritage Dictionary of the English Language*. Houghton Mifflin, New York.

Text Encoding Initiative. 1990. TEI P1 guidelines for the encoding and interchange of machine readable texts. Technical report, Text Encoding Initiative. Draft Version 1.1 1, `http://www.tei-c.org/Vault/Vault-GL.html`.