Linguistic Linked Open Data (LLOD) – Building the cloud

Christian Chiarcos
Goethe-University Frankfurt am Main, Germany
chiarcos@informatik.uni-frankfurt.de

The last decades have seen an immense maturation of Natural Language Processing (NLP) and an increased interest to apply NLP techniques and resources to real-world applications in business and academia. This process has certainly been facilitated by the increased availability of language data in the internet age, and the subsequent paradigm shift to statistical approaches, but also it coincided with an increasing acceptance of empirical approaches in linguistics and related academic fields, including empirical approaches to typology (Greenberg, 1963), corpus linguistics (Francis and Kucera, 1979, Brown Corpus), and (computational) lexicography (Kucera, 1969), as well as the dawn of Digital Humanities (Busa, 1974).

Given the complexity of language and the analysis of linguistic data on different levels, its investigation involves a broad band-width of formalisms and resources used to analyze, process and generate natural language. With the transition to empirical, data-driven research, the primary challenge in the field is thus to store, connect and exploit the wealth of language data available in all its heterogeneity. **Interoperability** of language resources has hence been an important issue addressed by the community since the late 1980s (Text Encoding Initiative, 1990), but still remains a problem that is solved only partially, i.e., on the level of specific sub-types of linguistic resources, such as lexical resources (Francopoulo et al., 2006) or annotated corpora (Ide and Suderman, 2007), respectively. A closely related challenge is **information integration**, i.e., how information from different sources can be retrieved and combined in an efficient way.

Recently, both challenges have been addressed by means of Linked Data principles (Chiarcos et al., 2013a,b), eventually leading to the formation of a **Linguistic Linked Open Data (LLOD) cloud** (Chiarcos et al., 2012b). The talk describes its current state of development, it presents selected examples for main types of linguistic resources in the LLOD cloud, and objectives leading to the adaptation of Linked Data principles for any of these.

Further, the talk elaborates on history and goals behind this effort, its relation to established standardization initiatives in the field, and on-going community activities conducted under the umbrella of the **Open Linguistics Working Group (OWLG)** of the Open Knowledge Foundation (Chiarcos et al., 2012a), an initiative of experts from various fields concerned with linguistic data, which works towards

1. promoting the idea of open linguistic resources,
2. developing means for their representation, and
3. encouraging the exchange of ideas and resources across different disciplines.

As the Linked Data paradigm can be used to facilitate any of these aspects, the OWLG identified potential application scenarios for linked and/or open resources in linguistics since its formation in 2010. The Working Group also has intensified its **community-building efforts** by means of a series of workshops, accompanying publications and data set releases. As a result of this process, numerous resources have been provided in a Linked-Data-compliant way, and linked with each other, as sketched here for selected examples.

References

Roberto Busa. *Index Thomisticus*. Frommann-Holzboog, Stuttgart, 1974.

Christian Chiarcos, Sebastian Hellmann, Sebastian Nordhoff, Steven Moran, Richard Littauer, Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian

Proceedings of the Joint Workshop on NLP&LOD and SWAIE, pages 1–2,
Hissar, Bulgaria, 12 September 2013.

M. Meyer. The Open Linguistics Working Group. In *Proc. 8th International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3603–3610, Istanbul, Turkey, May 2012a.

Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann. *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*. Springer, Heidelberg, 2012b.

Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum. Towards open data for linguistics: Lexical Linked Data. In Alessandro Oltramari, Piek Vossen, Lu Qin, and Eduard Hovy, editors, *New Trends of Research in Ontologies and Lexical Resources*. Springer, Heidelberg, 2013a.

Christian Chiarcos, Steven Moran, Pablo N. Mendes, Sebastian Nordhoff, and Richard Littauer. Building a Linked Open Data cloud of linguistic resources: Motivations and developments. In Iryna Gurevych and Jungi Kim, editors, *The People's Web Meets NLP. Collaboratively Constructed Language Resources*. Springer, Heidelberg, 2013b.

W. Nelson Francis and Henry Kucera. *Brown Corpus Manual. Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*. Providence, Rhode Island, 1979. URL `\url{http://icame.uib.no/brown/bcm.html}`. original edition 1964.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. Lexical Markup Framework (LMF). In *Proc. 5th International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, May 2006.

Joseph Greenberg. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph Greenberg, editor, *Universals of Language*, pages 58–90. MIT Press, Cambridge, 1963.

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

Henry Kucera. Computers in language analysis and lexicography. In *American Heritage Dictionary of the English Language*, pages XXXVII–XL. Houghton Mifflin, New York, 1969.

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