Corpora and Linguistic Linked Open Data: Motivations, Applications, Limitations

Christian Chiarcos[1]
(1) Applied Computational Linguistics
Johann Wolfgang Goethe Universität Frankfurt a. M.
60054 Frankfurt am Main, Germany
chiarcos@informatik.uni-frankfurt.de

Linguistic Linked Open Data (LLOD) is a technology and a movement in several disciplines working with language resources, including Natural Language Processing, general linguistics, computational lexicography and the localization industry. This talk describes basic principles of Linguistic Linked Open Data and their application to linguistically annotated corpora, it summarizes the current status of the Linguistic Linked Open Data cloud and gives an overview over selected LLOD vocabularies and their uses.

A resource constitutes Linguistic Linked Open Data if it is published in accordance with the following principles :

1. The dataset is relevant for linguistic research or NLP algorithms.
2. The elements in the dataset should be uniquely identified by means of a URI.
3. The URI should resolve, so users can access more information using web browsers.
4. Resolving an LLOD resource should return results using web standards such as Resource Description Framework (RDF).
5. Links to other resources should be included to help users discover new resources and provide semantics.
6. Data should be openly licensed using licenses such as the Creative Commons licenses.

Criterion (1) defines **linguistic**(ally relevant) data, criteria (2-5) define **linked** data, criterion (6) defines **open data**, their combination thus yields Linguistic Linked Open Data.

The primary benefits of LLOD have been identified as :

- Representation : Linked graphs are a more flexible representation format for linguistic data
- Interoperability : Common RDF models can easily be integrated
- Federation : Data from multiple sources can trivially be combined
- Ecosystem : Tools for RDF and linked data are widely available under open source licenses
- Expressivity : Existing vocabularies help express linguistic resources.
- Semantics : Common links express what you mean.
- Dynamicity : Web data can be continuously improved.

I specifically focus on linguistically annotated corpora and discuss the potential of Linked Data in relation to four standing problems in the field :

1. representing highly interlinked corpora (e.g., multi-layer corpora, annotated parallel corpora),

2. integrating corpora with lexical resources available from the web of data,
3. facilitating annotation interoperability using terminology resources available from the web of data, and
4. streamlining data manipulation processes in a modular and domain-independent fashion.

These aspects will be discussed in relation to two selected resources from both general linguistics and Natural Language Processing. Finally, the talk will discuss some of the challenges that LLOD is still facing in both areas.

Références

CHIARCOS C., HELLMANN S. & NORDHOFF S. (2011). Towards a linguistic linked open data cloud : The open linguistics working group. *Traitement automatique des langues*, **52**(3), 245–275.

CHIARCOS C., MCCRAE J., CIMIANO P. & FELLBAUM C. (2013). Towards open data for linguistics : Lexical linked data. In A. OLTRAMARI, P. VOSSEN, L. QIN & E. HOVY, Eds., *New Trends of Research in Ontologies and Lexical Resources*. Heidelberg : Springer.

CHIARCOS C. & SUKHAREVA M. (2015). OLiA - ontologies of linguistic annotation. *Semantic Web Journal*, **6**, 379–386.

CHIARCOS C. *et al.* (2016a). CoNLL-RDF. beyond the tsv. unpublished manuscript.

CHIARCOS C. *et al.* (2016b). Leight-weight conceptual interoperability for the universal dependencies. unpublished manuscript.

MCCRAE J. P., CHIARCOS C., BOND F., CIMIANO P., DECLERCK T., DE MELO G., GRACIA J., HELLMANN S., KLIMEK B., MORAN S., OSENOVA P., PAREJA-LORA A. & POOL J. (2016). The open linguistics working group : Developing the linguistic linked open data cloud. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, p. 2435–2441, Portorož, Slovenia : European Language Resources Association (ELRA).

SUKHAREVA M. & CHIARCOS C. (2016). Combining ontologies and neural networks for analyzing historical language varieties. a case study in middle low german. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, p. 1471–1480, Portorož, Slovenia : European Language Resources Association (ELRA).