Introduction to Linguistic Linked Open Data

Christian Chiarcos
Goethe University Frankfurt, Germany

Abstract

The number of resources that provide lexical data keeps continuously increasing and quantity and diversity, as a result of academic research in (computational) linguistics, digital humanities, and e-lexicography, but also as a requirement of or components for applications of language technologies in industry and academia. This vast landscape of heterogeneous and often isolated language resources creates obstacles for their straightforward linking and integration in data processing pipelines in an interoperable manner. To address this, experts working at the intersection of natural language processing, knowledge representation (Semantic Web) and the language sciences have adopted approaches to linguistic data representation based on the Linked Open Data (LOD) paradigm, giving birth to the Linguistic Linked Open Data (LLOD) cloud and designated LLOD technologies. In this context, linked data emerges as a way to make linguistic data uniformly query-able, interoperable, and easily discoverable as well as reusable on the basis of web standards. This tutorial will provide attendees a theoretical and practical overview of the foundations of LLOD, covering, among other aspects, an introduction to the Semantic Web and linked data, and a walkthrough of the different steps for linguistic linked data generation. We will lay special emphasis on knowledge representation with the de-facto standard for lexical data representation on the Web, the OntoLex-Lemon model, and other linguistic vocabularies, and using such data for performing cross-lingual search in the web of data.

Biography. Christian Chiarcos is a computer scientist and linguist with a specialization in the processing of heterogeneous linguistic data. Following studies at Technical University Berlin, Humboldt University Berlin and University of Potsdam, Germany, he received a PhD in Computational Linguistics in 2010, with a thesis on anaphora and information structure in the context of natural language generation. Besides a general research focus on computational semantics and discourse, he has worked on matters of interoperability in natural language processing, linguistics and the philologies since 2005, and has subsequently become an expert on creating, maintaining, processing and consolidating linguistic data. Following his PhD studies, he joined the Information Sciences Institute of the University of Southern California. From 2013 to 2022, he has been Professor (W1) at the Institute for Computer Science at Goethe University Frankfurt, Germany, and heading the Applied Computational Linguistics (ACoLi) Lab. He has been leading the Early Carreer Research Group "LiODi. Linked Open Dictionaries" funded by the German Federal Ministry of Education and Research (BMBF, 2015-2022) and has been active in

a large number of German, European and international research projects. In May 2022, he joined the Cologne Center for eHumanities (CCeH) and is currently working at the Institute for Digital Humanities at the University of Cologne, Germany. Aside from research interests in computational semantics and language technology, his recent activities include applications of these technologies in the language sciences, in the industry and in the humanities, with notable results such as the first syntactic parser for medieval German (Middle High German, 2018), the publication of the first machine-generated science book (2019: Lithium-Ion Batteries. A Machine-Generated Summary of Current Research, Springer, Cham), and the first machine translation system for Sumerian cuneiform (2020).