Developing ready-to-use lexical knowledge graphs for a multilingual next generation internet

Knowledge processing over different languages and information integration over distributed data represent the fundamental cornerstones of existing and emerging infrastructures and services that constitute, for example, the Digital Single Market formulated as a strategic goal by the European Commission. Yet, both aspects, multilinguality and heterogeneity of resources and tools, involve a number of technical challenges addressed by recent efforts to develop ready-to-use technologies for a multilingual next generation internet.

In this talk, we address the challenge to harmonize existing lexical data into a coherent and interlinked representation: Machine-readable dictionaries come in various shapes, in different formats, with different kinds of markup and different metadata. We describe an on-going effort to consolidate such data into a massive, multilingual knowledge graph on the basis of RDF and the OntoLex-Lemon vocabulary. We put a particular focus on two aspects of this effort, the transformation of heterogeneous dictionaries into a coherent form, and the creation of inter-dictionary links. So far, we produced a lexical knowledge graph covering more than 500 languages and data from more than 3000 dictionaries, parts of this data are available via our GitHub repository.

A major component of this workflow is the Flexible Integrated Transformation and Annotation eNgineering (FINTAN) platform, a modular, generic and extensible system for developing complex transformation and information enrichment pipelines with workflow management and visualization features for the transformation and enrichment of natural language data and annotations, e.g., vocabulary normalization and entity linking. We illustrate the application of FINTAN technology for the example of converting and consolidating the Apertium dictionaries, a large-scale collection of bilingual dictionaries, and their linking with ontologies of linguistic terminology, lexinfo and OLiA.

Prof. Dr. Christian Chiarcos
Goethe-Universität Frankfurt
room G 300, Fr. 13.12.2019, 13:00