Linked Data as a Cornerstone of Linguistic Data Science

Jorge Gracia, University of Zaragoza, Spain, jogracia@unizar.es

Abstract

At present, we are witnessing incredible advancements in language technologies, natural language processing, and related fields, mostly stimulated by the success of deep learning technologies and the increasing availability of huge amounts of textual data on the Web. Algorithms and models that were commonly used a few years ago are rapidly substituted by newer, more effective ones, with little time for researchers and practitioners to adapt to them. In such an evolving and challenging scenario, what is the role of linguistic data science? We understand linguistic data science as a subfield of data science which focuses on the systematic analysis and study of the structure and properties of linguistic data at a large scale, along with methods and techniques to extract new knowledge and insights from it. To that end, it is necessary to provide a formal basis to the analysis, representation, integration, and exploitation of such data at their different levels (syntax, morphology, lexicon, etc.).

In this talk, we will try to answer the previous question, for which we first have to consider the difference between textual data, where current trends put the focus, and linguistic data (with their linguistic features explicitly represented). A reflection on the type of problems that can be solved with the use of the latter type of data is needed. We will also discuss how linked data technologies can be essential to provide a basis for linguistic data science, by enabling an ecosystem of multilingual and semantically interoperable linguistic data at Web scale. We will review recent advancements and open challenges in the field of linguistic linked data, mentioning recent studies carried out in the context of the NexusLinguarum COST Action on that matter. Among other things, they identify the main current challenges that the linguistic linked data community needs to face to maximize the adoption of their technologies. These include the need to break some entry barriers to enable non-experts to adopt the technology seamlessly, the need for more sustainable hosting solutions, as well as developments targeted to increase multilinguality and support for under-resourced languages on the Web. Finally, some examples of how linguistic linked data technologies are currently being exploited for research in linguistics and industrial applications will be provided.

Keywords: linguistic data science, linked data, linguistic linked data