Towards an open ecosystem of multilingual interoperable linguistic data

A Policy Brief about the social and technological interest of Linguistic Data Science

Authors: Thierry Declerck, Jorge Gracia, John McCrae, Philipp Cimiano, Paul Buitelaar

Overcoming language barriers in Europe and worldwide is a challenge that is addressed in different ways. One approach consists of developing powerful machine translation systems, which are helping the different administrations to communicate among each other and with citizens, offering access to relevant data in their own language and across languages. For instance, the European eTranslation infrastructure has been put in place and is being further developed for reaching this goal.[1] However, such machine translation engines need to be "fed" with the right type of language data, like parallel corpora, terminologies, etc, in the right format and covering the right domain of applications, like eHealth, eJustice, etc.

This need can in fact be addressed by another and complementary approach for overcoming language barriers, which is based on the construction of a mature holistic ecosystem of multilingual and semantically interoperable linguistic data. This is the main challenge addressed by the NexusLinguarum COST Action "CA18209 - European network for Web-centred linguistic data science". Such an ecosystem provides for in-depth description of language data, so that they can be optimally offered for different multilingual applications, including but not limited to, machine translation, for example, in an integrated European lexicography infrastructure[2], or a harmonized European multilingual terminology database,[3] which can be extended to domains beyond EU-specific terminology. Multilingual language data are also at the core of many academic and teaching institutions, where they need to be encoded to allow scholars to immediately start their work in the fields of Digital Humanities, Sociology, Political Sciences, Medicine, Law, etc. Dealing with language data in such a way that they can be made available to everyone, in science, economy, medicine, culture or in governments is the object of what we call (open) "linguistic data science".

We understand linguistic data science as a subfield of the "data science" field, which focuses on the systematic analysis and study of the structure and properties of data at a large scale, along with methods and techniques to extract new knowledge and insights from it. Linguistic data science is a specific case of data science, which is concerned with providing a formal basis for the analysis, representation, integration, and exploitation of all types of language data. Linguistic data go from the lexical and morphological levels, to syntactic and semantic levels of analysis.

Psycho- and neurolinguistics are studying mental and brain activities, in which language related information can be detected on the base of neuronal encodings that do not necessarily correspond to

---

[1] See https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en

[2] This goal is pursued by the European H2020 project ELEXIS (European Lexicography Infrastructure). See https://elex.is/ for more details

[3] See the *IATE* (Interactive Terminology for Europe) is the EU's terminology database, which "It has been used in the EU institutions and agencies since summer 2004 for the collection, dissemination and management of EU-specific terminology." (https://iate.europa.eu/about)

the language data we are confronted with in written documents. The efficient processing of huge text and also speech and even video corpora for, e.g. training machine translation systems, speech analysers and synthesizers, sign languages analysers, requires the transformation of human-readable language data into machine processable formal representations, resulting in vector spaces, in which language data can be encoded for developing powerful NLP applications. As a consequence, besides linguistic knowledge, also strong skills in mathematics are needed in language technology. A cross-disciplinary is, therefore, paramount in the field of linguistic data science.

To respond to the above mentioned challenge, we need to establish synergies across Europe between linguists, computer scientists, terminologists, language professionals, and other stakeholders in industry and society, in order to establish the area of linguistic data science. We need to design and implement use cases leading to an integrated ecosystem in which different types of language data will be made available in a standardized way, using instruments and open standards introduced by the W3C as the means for ensuring intelligent access, integration and distribution of language data tailored to the needs of European citizens, scientists, administrations and companies.

Such an ecosystem, unavailable today, is key in order to foster the systematic cross-lingual discovery, exploration, exploitation, extension, curation, and quality control of linguistic data. Linked data (LD) technologies, in combination with natural language processing (NLP) techniques and multilingual language resources (LRs) (bilingual dictionaries, multilingual corpora, terminologies, etc.), have the potential to enable such an ecosystem that will allow for transparent information flow across linguistic data sources in multiple languages.

Such an ecosystemThe proposed framework has important economic implications, as it facilitates the commercial exploitation of linguistic data, reducing costs concerning (re)-use of existing data by following interoperable standards for data representation and rich metadata descriptions for making the data findable. It further reduces the uncertainty of using data by making explicit the conditions under which the data can be used by appropriate licensing information.

Our goal is to build on proofs of concepts developed in the context of H2020 projects such as Prêt-à-LLOD[4] and others as a basis to describe successful best practices of how interoperable linguistic linked data can strengthen the competitive advantage of the European economy in innovative fields such as digitized healthcare, open government and electronic lexicography.

NexusLinguarum aims to develop a common understanding, standards, and best practices in the field of language data science for supporting the Digital Single Market, cross-border commerce, cultural exchange, and communication in Europe.

Acknowledgements: We thank Ilan Kernerman, Julia Bosque-Gil, Sara Carvalho, Penny Labropoulou, and Marieke van Erp for their careful review and valuable comments.

[4] See https://pret-a-llod.github.io/