

[Click here to be redirected to the virtual room of the Project Expo.](#)

Project abbreviation: NexusLinguarum



Project name: European network for Web-centred linguistic data science [COST Action CA18209]

Project coordinator: Universidad Zaragoza, Spain

Project consortium: 36/38 COST Members, 1 Cooperating Member, 1 Specific organization, 3 NNC and 2 IPC Countries

Funding: European Cooperation in Science and Technology (COST)

Project duration: 28/10/2019 - 27/10/2023

Main key words: linguistic data science, multilingualism, linguistic linked data, language resources

Background of the research topic: In recent years, various efforts have arisen with regard to the representation and publication of linguistic resources such as lexicons, dictionaries, corpora, terminologies and linguistic ontologies. These efforts have exploited Semantic Web technologies and the Linguistic Linked Data (LLD) publication paradigm to facilitate and enhance the discovery, interoperability, integration and reusability of language resources. Initiatives such as the H2020 projects ELEXIS and Prêt-à-LLOD and the COST Action NexusLinguarum aim at developing robust ecosystems and networks of experts to address the LLD lifecycle, from identifying the requirements concerning the representation of linguistic resources to their exploitation by natural language processing (NLP) applications. With the rapid growth of the Linguistic Linked Open Data (LLOD) cloud and the increasing interest in the use of linked data for NLP, new challenges emerge concerning particular use cases and domain applications, language-specific features and quality dimensions, the evolution of LLD resources throughout time and the leverage of linguistic resources along LD technologies in NLP research, among other diverse aspects.

Goal of the project: The construction of a mature holistic ecosystem of multilingual and semantically interoperable linguistic data is required at Web scale. Such an ecosystem, unavailable today, is needed to foster the systematic cross-lingual discovery, exploration, exploitation, extension, curation and quality control of linguistic data. We argue that 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, by addressing the semantic interoperability problem.

Project abstract: The main aim of this Action is to promote synergies across Europe between linguists, computer scientists, terminologists, and other stakeholders in industry and society, in order to investigate and extend the area of linguistic data science. We understand linguistic data science as a subfield of the emerging “data science”, 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, which is concerned with providing a formal basis to the analysis, representation, integration and exploitation of language data (syntax, morphology, lexicon, etc.). In fact, the specificities of linguistic data are an aspect largely unexplored so far in a big data context.

Publications:

Recent Developments for the Linguistic Linked Open Data Infrastructure. 12th Conference on Language Resources and Evaluation (LREC 2020)

[Click here to be redirected to the virtual room of the Project Expo.](#)

COST Action "European network for Web-centred linguistic data science" (NexusLinguarum).
Procesamiento del Lenguaje Natural (2020)

Special Issue on Latest Advancements in Linguistic Linked Data. Semantic Web Journal (to appear
2022)