The Sixth *NexusLinguarum* Plenary and Final Meeting (March 20-21, 2024, Athens)

PAPER SUBMITTED:
28 Jun 2024

PAPER ACCEPTED:
29 Jun 2024

Ranka Stanković
ranka@rgf.rs

ORCID: 0000-0001-5123-6273

University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

The COST action *CA18209*[1] - *European network for Web-centered linguistic data science* (*NexusLinguarum*[2]) started in October 2019 and ended in April 2024. *NexusLinguarum* Management Committee comprised 69 members from 38 countries. The 213 participants from 39 countries were members of working groups (WGs). The main aim of the Action was to promote synergies across Europe between linguists, computer scientists, terminologists, and other stakeholders in industry and society, to analyze and extend the area of linguistic data science. The linked data (LD) technologies, in combination with natural language processing (NLP) techniques and multilingual language resources (LRs) (bilingual dictionaries, multilingual corpora, terminologies, etc.), were considered as the potential to provide an ecosystem that will allow for transparent information flow across linguistic data sources in multiple languages, by addressing the semantic interoperability problem.

The sixth plenary meeting of the COST action *CA18209 - European network for Web-centered linguistic data science* (*NexusLinguarum*) took place in Athens on the 20th and 21st March 2024. It was a hybrid meeting, followed by a series of additional events due to the end of the Action, such as the recording of lessons that will be used for a MOOC (*Massive Open Online Course*), and a poster session.

The first day was dedicated to the achievements, deliverables, and outcomes. The reports are given for each working group (WG). The results of WG1: *Linked data-based language resources* were presented by Dr. Milan Dojchinovski. Patricia Martín Chozas reported on achievements of WG2:

[1] CA18209
[2] *NexusLinguarum*

Linked data-aware NLP services. Dagmar Gromann presented activities in WG3: *Support for Linguistic Data Science*, while the overview of WG4: *Use Cases and Applications was given by Sara Carvalho.*

During the presentation of *NexusLinguarum* main results, enlightening talks and posters were offered. Maria Pia di Buono talked about leveraging Large Language Models (LLMs) for Linguistic Linked Data (LLD). Maciej Ogrodniczuk presented a project proposal on *Universal Discourse.* Linguistic Linked Open Data (LLOD) for interoperable morphological description was the topic of Max Ionov, while Ineke Schuurman reported on sign languages and LLOD, Deliverable 4.3 with Final Activity Report with use cases and applications (Carvalho and Kernerman 2024) offered a comprehensive description of the nine use cases that were realized in the last reporting period, and their implementation. The use cases and applications were used for testing and validating the Action's relevant methodologies, technologies, and standards. Use cases on public health were reported by Ana Ostroški Anić. The offensive Language categorization gold standard for the LLOD schema was presented by Anna Bączkowska. Florentina Armaselu described the use case LLODIA (*LLOD for Diachronic Analysis*), and Giedre Valunaite Oleskevicienė presented the use case in Social Sciences. The deep learning for linguistic data analysis was surveyed by Atanas Hristov. Hugo Gonçalo Oliveira presented *MultiLexBATS: Description and Analogy Completion* (Gromann et al. 2024). Paola Marongiu elaborated on lexical semantic change detection in Latin with an use-case on medical Latin.

These enlightening talks were followed by the poster session, where Radovan Garabik presented experiences on using LLOD to bootstrap bilingual dictionaries. Verginica Barbu Mititelu gave an overview of Romanian LD resources developed during *NexusLinguarum*, Ranka Stanković recapitulated the approach and experience on linking a NIF (*NLP Interchange Format*) corpus to an *Ontolex-Lemo* dictionary on Wikibase. Katerina Zdravkova introduced the results on resolving inflectional ambiguity of Macedonian adjectives, while Dimitar Trajanov presented the approach for automating the process of dictionary creation.

Some of the results of short-term scientific missions (STSM) and virtual mobility grants were also presented. The use case on cybersecurity was presented by Sigita Rackevičienė, who outlined the development of cybersecurity language resources. The second topic related to cybersecurity was introduced by Christian Chiarcos, related to converting and linking cybersecurity linguistic datasets. A new *UD-Treebank* for the Albanian Language was summarized by Manjola Zacellari. Anas Fahad Khan reported on the

STSM on Domain Labels in Linked Lexicographic Resources. Aleksandra Tomaszewska recapitulated the *Multilingual Discourse Annotation Initiative (MDAI): Objectives, Challenges, and Possibilities.*

On the second day, Rute Costa outlined the *Academic Common Curriculum on Linguistic Data Science* as a collaborative effort to craft a proposal for an *Erasmus Mundus Joint Master* to the European Union. The first stage involved the development of the *Linking Linguistics to Data Science (LL2DS)* project, aiming at establishing a groundbreaking *Erasmus Mundus Joint Master*, which would be recognized globally for its innovative fusion of Linguistics and Data Science. This project led to the second phase, which involved submitting a fully designed EMJM (Costa and Garcia 2024).

A roadmap with a common agenda for future research on linguistic data science was presented. Its first part identifies and discusses a series of challenges in the field of Linguistic Data Science (LDS), more particularly in LLOD. The main issues were entry barriers to the technology, sustainability, coverage of current representation models, metadata, cross-lingual linking, under-resourced languages, and multilinguality. The relation between LLOD and the emergent LLMs was especially highlighted, and a concrete plan to continue the activities of the *NexusLinguarum COST Action*, by continuing to progress along the proposed roadmap, was provided.

Preparation of *Guidelines and Best Practices on LLOD* (Martín Chozas et al. 2024) is a collaborative effort of the *NexusLinguarum* COST Action and the Best Practices for Multilingual Linked Open Data (BPMLOD) W3C Community Group to develop guidelines and best practices for linking data and services across languages. The first set of guidelines focuses on LLOD, and the following, on integrating LLOD with Natural Language Processing (NLP) services. The comprehensive recommendations for LLOD and LLOD-aware NLP services, together with the evolution of the BPMLOD community and ongoing efforts to update guidelines, can ensure sustainability beyond the *NexusLinguarum* COST Action.

Acknowledgment

This work has been supported by the COST action CA18209 - European network for Web-centred linguistic data science (NexusLinguarum).

References

Carvalho, Sara, and Ilan Kernerman. 2024. *Final Activity Report (months 25-54). Working Group 4: Use Cases and Applications.* https://nexuslinguarum.eu/wp-content/uploads/2024/04/Deliverable-D4.3-WG4-Final-Activity-Report_compressed.pdf.

Costa, Rute, and Jorge Garcia. 2024. *Academic Common Curriculum on Linguistic Data Science - LDS.* https://nexuslinguarum.eu/wp-content/uploads/2024/04/Deliverable-D3.3-common-curriculum-1.pdf.

Gromann, Dagmar, Hugo Gonçalo Oliveira, Lucia Pitarch, and et al. 2024. "MultiLexBATS: Multilingual Dataset of Lexical Semantic Relations." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024),* 11783–11793. ELRA Language Resource Association. https://aclanthology.org/2024.lrec-main.1029.pdf.

Martín Chozas, Patricia, Milan Dojchinovski, Katerina Gkirtzou, Anas Fahad Khan, and Andon Tchechmedjiev. 2024. *Guidelines and Best Practices on Linguistic Linked Open Data.* https://nexuslinguarum.eu/wp-content/uploads/2024/03/Deliverable-D1.4-Guidelines-and-Best-Practices-on-LLOD.pdf.