NexusLinguarum: European network for Web-centred linguistic data science

Jorge Gracia

The kickoff meeting of the newly created 'European network for Web-centred linguistic data science' (*NexusLinguarum* in its short name) took place in Brussels, Belgium, on 28 October 2019. The meeting brought together representatives of the 33 countries constituting the initial network to discuss the objectives, the plans for implementing the different networking tools, and the scope and goals of the different working groups, as well as to elect the action's management board. The initial group consisted of a broad network of experts from different areas, like computer science, semantic web, artificial intelligence, linguistics, humanities, etc. That was the beginning of an exciting journey towards building a common ecosystem to support research on *linguistic data science* in a Web-centred context.

We understand linguistic data science as a subfield of the growing data science field that 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. Linguistic data science is concerned with providing a formal basis to the analysis, representation, integration and exploitation of linguistic data for language analysis (e.g. syntax, morphology, terminology, etc.) and language applications (e.g. machine translation, speech recognition, sentiment analysis, etc.).

NexusLinguarum will last for four years and is funded by the European Cooperation in Science and Technology (COST) organization, which supports such highly competitive projects (COST Actions) by financing research networks on emerging issues, through mechanisms such as research visits, organization of congresses, scientific meetings, summer schools, etc.

To enable the study of linguistic data in the most productive and efficient ways, the NexusLinguarum COST Action is set to enhance the construction of an ecosystem of multilingual and semantically interoperable linguistic data at the scale of the Web. To this end, methods and techniques of the Semantic Web, Natural Language Processing and Language Resources are studied and combined. Such an ecosystem could reduce language barriers in Europe (and eventually beyond) and favour both electronic commerce and cultural

Jorge Gracia is the Chair of the NexusLinguarum 'European network for Web-centred linguistic data science' COST Action. He works as assistant professor at the Department of Computer Science and Systems Engineering (University of Zaragoza, Spain) as a member of the Aragon Institute of Engineering Research (I3A) and of the Distributed Information Systems research group. His main research interests are Semantic Web, Ontology Matching, Multilingual Web of Data, Query Interpretation, and Web Intelligence, and his recent work focuses on linked data-based lexicography as well as on methods and techniques for cross-lingual linking and information access. http://jogracia.url.ph/web/

Participants at NexusLinguarum kickoff meeting, Brussels, 28 October 2019

exchange between countries with different languages. Another objective is to support minority languages whose technological support is currently limited.

Through the study of Web-centred linguistic data science, we will be able to better understand the nature of language, through innovative methods for the representation, integration and comparison of linguistic data. Furthermore, since language is the medium in which human knowledge is transmitted, this field has the potential to decisively influence studies that use natural language for knowledge sharing, as is the case of the humanities, the legal domain, journalism, social sciences, etc.

Some of the main research coordination objectives of NexusLinguarum are to:

- propose, agree upon and disseminate best practices and standards for linking data and services across languages;
- organise activities to foster collaboration and communication across communities, such as scientific workshops involving broader communities to reach agreement on best practices;
- collect and analyse relevant use cases for linguistic data science and develop prototypes and demonstrators that will address some prototypical cases.

Furthermore, we plan to work out a curriculum for a Europe-wide master degree that the participating institutions could adopt to train a new generation of researchers in the area, thus introducing linguistic data science in a cross-discipline academic infrastructure.

Currently we count on participants from 42 countries (37 COST

Countries, 3 Near Neighbour Countries, and 2 International Partner Countries). So far, 137 members have joined the different working groups (WGs), a number which is steadily growing since the network is still open to new participants.

NexusLinguarum is organised in five working groups, four technical ones and a one for management activities:

WG1 – Linked data-based language resources. This WG lays the foundations to develop best practices for the evolution, creation, improvement, diagnosis, repair and enrichment of linguistic linked open data (LLOD) resources and value chains.

WG2 – Linked data-aware NLP services. This WG focuses on the application of linguistic data science methods including linked data to enrich NLP tasks in order to take advantage of the growing amount of linguistic (open) data available on the Web.

WG3 – Support for linguistic data science. This WG aims to foster the study of linguistic data by following data analytic techniques at a large scale in combination with LLOD and linked data-aware NLP techniques

WG4 – Use cases and applications. This WG focuses on studying use cases and practical applications of the relevant technologies involved in the Action.

WG5 – Management and dissemination. This WG takes care of the measures to be taken to ensure the creation of added value of the Action as a whole, to ensure its maximum visibility, and to monitor the cross-WG activities.

All these WGs have already started their activities, although they are still in initial phases. One of the fist outcomes to be delivered by NexusLinguarum is a study of use case definitions, which is currently under development by WG4. The following use cases are being analysed currently: Humanities and Social Sciences, Linguistics (Media and Social Media, and Language Acquisition), Life Sciences, and Technology (Cybersecurity and FinTech). The idea is to analyse the current state-of-the-art on each of these topics, analyse their needs and challenges, and determine the techniques and ideas of linguistic data science that might improve them, in close collaboration with the other WGs.

The other three technical WGs are also conducting initial surveys and analysing related projects and initiatives to set the ground for further development. Collaboration with other projects and initiatives are already on course, for instance with the W3C Linked Data for

Language Technologies community group in relation to the ongoing discussion towards a consolidated Linked Open Data vocabulary for linguistic annotations, in the context of WG1.

NexusLinguarum has already organised two face-to-face meetings of its Management Committee (MC): in Brussels in October 2019 (kickoff meeting), and in Prague (Czech Republic) in January 2020 collocated with the first WG meetings. The next MC + WGs meeting is due to take place in October 2020 in Lisbon (Portugal). In addition, regular teleconferences take place to enable and monitor the scientific progresses of the different WGs, and a number of training schools and scientific events are planned for 2021.

More information can be found at https://nexuslinguarum.eu/. New participants can join the network through this registration form https://forms.gle/ZML87XLHnxXdPrbh6.

NexusLinguarum Core group

Chair. Jorge Gracia
University of Zaragoza, Spain

Vice-chair. John McCrae
National University of Ireland, Galway, Ireland

Grant Holder Scientific representative. Elena Montiel-Ponsoda
Universidad Politécnica de Madrid, Spain

Science Communication manager. Thierry Declerck
DFKI, Germany

Short Term Scientific Missions coordinator. Penny Labropoulou
Athena Research Center, Greece

Inclusiveness Target Countries Conference Grant coordinator. Vojtech Svatek
University of Economics, Prague, Czech Republic

WG1 leader. Milan Dojchinovski
Czech Technical University in Prague, Czech Republic

WG1 co-leader. Julia Bosque-Gil
University of Zaragoza, Spain

WG2 leader. Marieke van Erp
KNAW Humanities Cluster, The Netherlands

WG3 leader. Dagmar Gromann
University of Vienna, Austria

WG3 co-leader. Amaryllis Mavragani
University of Stirling, UK

WG4 leader. Sara Carvalho
University of Aveiro, Portugal

WG4 co-leader. Ilan Kernerman
K Dictionaries, Israel

An overview of NexusLinguarum Working Groups

The Action is composed of five working groups (WGs) interoperating and providing mutual feedback. They cover, in a bottom-up approach, the technical and infrastructural groundings needed to attain the objectives of the Action along with a range of use cases and applications. In addition to their own tasks, all WGs participate in preparing cross-group dissemination activities. The scientific work is carried out over four years through workshops and other meetings as well as remote cooperation through electronic communication means (email, teleconference, etc).

* Short-Term Scientific Missions (STSMs) organized within each WG, to promote synergies and maximize cooperation, and International Training Schools (ITCs), are not included in this overview.

WG1 – Linked data-based language resources

Objective. Lay the foundations and develop best practices for the evolution, creation, improvement, diagnosis, repair and enrichment of linguistic linked open data (LLOD) resources and value chains.

Tasks

Task 1.1: LLOD modelling. Update, extension and improvement of existing models for representing linguistic information as linked data (LD, e.g. lemon-ontolex, LexInfo, OLiA, NIF, etc).

Task 1.2: Creation and evolution of LLOD resources in a distributed and collaborative setting. Analysis of new approaches for the distribution and collaborative creation and extension of linguistic resources to facilitate the extension of existing resources and their publication as LD.

Task 1.3: Cross-lingual data interlinking, access and retrieval in LLOD. Studying novel (semi-)automatic methods aimed to increase the interlinking across LLOD datasets, and methods and techniques for accessing and exploiting data on the Web across different languages, based on the use of linguistic linked data (LLD).

Task 1.4: Improving and monitoring quality of LLOD sources. New techniques to monitor and improve the quality of LLOD sources by novel approaches for diagnosis and repair and new measures allowing to monitor and assess the quality of such sources, as well as analyzing semi-automatic and automatic methods for validating LD and cross-resource links via collaborative strategies.

Task 1.5: Development of the LLOD cloud for under-resourced languages and domains. Analysis and development of language technologies serving under-resourced languages and domains in the LLOD cloud.

Deliverables

- Scientific papers on linguistic linked data and language resources
- Training school on linguistic linked data
- Guidelines and best practices on the generation, interlinking, publication and validation of LLOD (new and update of existing ones)
- Policy brief about the inclusion of data from under-resourced languages
- Intermediate and final activity reports

WG2 – LD-aware Natural Language Processing services

Objective. Applying LLD to enrich NLP tasks taking advantage of the growing amount of linguistic linked (open) data available on the Web.

Tasks

Task 2.1: LLD in Knowledge Extraction. Analysis of large-scale integrated linguistic and semantic knowledge for multiple domains and languages to open up new possibilities in taxonomy and ontology-based information extraction.

Task 2.2: LLD in Machine Translation. Incorporating multilingual LLD in machine translation (MT), both syntactically (e.g. using dependency relations) and semantically (using lexical semantics), as well as exploring LD for expressing translation workflow metadata to improve MT output.

Task 2.3: LLD in Multilingual Question Answering. Examining how lexical knowledge required by QA systems can be extracted from LLD.

Task 2.4: LLD in Word Sense Disambiguation and Entity Linking. Studying the impact of LLD on disambiguation in multilingual content processing, such as for the translation of terms and idioms in user-generated content to detect words or phrases used in a potentially offensive manner.

Task 2.5: LLD in Terminology and Knowledge Management. Cross-disciplinary research on applying LLD in multilingual terminology and knowledge resource management, including their linking, merging with enterprise (proprietary) resources, and publishing on the Web as part of a global ecosystem of multilingual data.

Deliverables

- Scientific papers on linked data-aware NLP
- Guidelines and best practices on LLOD and NLP (new and update of existing ones)
- Intermediate and final activity reports

WG3 – Support for linguistic data science

Objective. Understand linguistic data by following data analytic techniques at a large scale in combination with LLOD and LD-aware NLP techniques, covering scalability issues in the study of multilingual linguistic data given the fact that datasets are rapidly growing in size, leading to huge amounts of data on the Web (*big data*).

Tasks

Task 3.1: Big data and linguistic information. Studying big data sources and state of the art statistical analysis in combination with LLOD to better understand language, also considering visual analytics, having an impact on the linguistics aspect in all sub-domains, from typology to syntax to comparative linguistics.

Task 3.2: Deep learning and neural approaches for linguistic data. Study the effective use of deep learning in understanding the specificities of linguistic data in a big data context, to be better exploited and combined with LD mechanisms.

Task 3.3: Linking structured multilingual language data across linguistic description levels. Explore how diverse data regarding phonology, morphology and lexicon that is spread across datasets of varying extent, quality and format, can be described, stored and accessed uniformly.

Task 3.4: Multidimensional linguistic data. Link language resources across various dimensions (such as time axis, style, genre, media, etc) to facilitate diachronic and sociolinguistic interoperable research.

Task 3.5: Education in linked data science. Develop a curriculum for linguistic data science in a cross-discipline academic infrastructure for a Europe-wide master's degree for training a new generation of researchers.

Deliverables

- Scientific papers on techniques that support linguistic data science
- Academic curriculum for the studies on linguistic data science
- Training school on linguistic data science
- Intermediate and final activity reports

WG4 – Use cases and applications

Objective. Exploring practical use cases and applications of the relevant methodologies and technologies involved in the Action.

Tasks

Task 4.1: Use cases in legal domain. Explore legal terminology and translation, and identify use cases ranging from unique identification and re-use of licenses at a Web-scale to assisted translation based on semantic annotations.

Task 4.2: Use cases in humanities and social sciences. Study how linguistic data science can deeply influence studies in the humanities and social sciences, allowing us to trace the history of the peoples of the world, understand literature and culture in new ways. and predict and analyze social trends.

UC4.2.1 – Use Case in Humanities

UC4.2.2 – Use Case in Social Sciences

Task 4.3: Use cases in linguistics. Investigate how linguistic data science and richer understanding of language can benefit research in linguistics (lexicography, typology, syntax, comparative linguistics, etc).

UC4.3.1 – Use Case in Media and Social Media

UC4.3.2 – Use Case in Language Acquisition

Task 4.4: Use cases in life sciences. Exploit structured linguistic information in the process of discovering hidden facts out of textual data, expanding text analytics techniques applied in biomedicine and other life sciences.

Task 4.5: Use cases in technology. Look into incorporating text analytics into advanced technological systems, such as for sentiment analysis and fake news detection, by developing customized domain-specific models.

UC4.5.1 – Use Case in Cybersecurity

UC4.5.2 – Use Case in Fintech

Deliverables

- Scientific papers on in-use applications of LLOD, NLP and linguistic big data
- Report describing requirements elicitation and use cases definitions
- Intermediate and final activity reports

WG 5 - Management and dissemination

Objective. Manage the measures taken to ensure the creation of added value of the Action as a whole and its optimal visibility, and monitor the cross-WG activities.

Tasks

Task 5.1: Management. Day-to-day management and administrative coordination.

Task 5.2: Cross-working group communication. Monitor communication across the different WG activities, especially in the MC and SC meetings.

Task 5.3: Capacity building. Coordinate the Short-Term Scientific Missions (STSMs) and elaborate a plan for the training schools, datathons and hackathons to be developed by the different WGs.

Task 5.4: External communication. Coordinate the Action's external communication including its website, social media streams, press releases, and the organization of (and participation in) events and workshops, etc.

Task 5.5: Scientific publications strategy. Monitor the scholarly publications produced in the Action, define a strategy for journal special issues, and provide the means to document them in a central repository (e.g. Zenodo).

Deliverables

- Roadmap document containing a common research agenda for linguistic data science
- Generated dissemination materials (blog entries, short reports of the different Action's events, press releases, etc.)
- Policy brief about the social and technological interest of linguistic data science