Nexus Linguarum

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

D3.4

Final Activity Report. Working Group 3 “Support for Linguistic Data Science ”

Main authors: Dagmar Gromann, Dimitar Trajanov, Radovan Garabík, Ineke Schuurman, Rute Costa

19 April 2024

Project Acronym NexusLinguarum

Project Title	European network for Web-centred linguistic data science
WG3 Title	Support for linguistic data science
COST Action	CA18029
Project Title	European network for Web-centred linguistic data science
Project Acronym	NexusLinguarum
Duration	54 months
Start of CA18029	October 2019
Project Website	https://nexuslinguarum.eu
Responsible Authors	Dagmar Gromann, Dimitar Trajanov, Radovan Garabík, Ineke Schuurman, Rute Costa
Contributors	Dagmar Gromann, Dimitar Trajanov, Radovan Garabík, Ineke Schuurman, Rute Costa, Ciprian-Octavian Truica
Reviewer	NexusLinguarum core group team
Version | Status	v1.0 | final
Date	19 April 2024

Acronyms List

CA Cost Action

DL Deep Learning

LD Linked Data

LLD Linguistic Linked Data

LLOD Linguistic Linked Open Data

LOD Linked Open Data

NLP Natural Language Processing

MC Management Committee

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses

UC Use Case

WG Working Group

Table of Contents

1. Executive Summary

Working Group 3 (WG3) of the NexusLinguarum COST Action entitled ***Support for linguistic data science*** aims to foster the study of linguistic data by following data analytic techniques at a large scale in combination with LLD and linked data-aware NLP techniques. These techniques range from **Big Data** and **deep learning** to different linguistic description levels for **multilingual** and **multimodal** representations. Additionally, **education** in linguistic data science is one dedicated task of this WG.

This deliverable reports on all activities of WG3 during the second half of the COST Action NexusLinguarum, including meetings, organized events, surveys, publications, and collaborations with other working groups as well as other initiatives and communities. Activities of this working group related to the first half of the COST Action NexusLinguarum can be found in Deliverable D3.1 submitted in October 2021. Furthermore, this report details planned initiatives and steps taken to continue the work of individual tasks and keep the community effort active after the end of NexusLinguarum.

2. Introduction

Support for linguistic data spans from investigating Big Data and deep learning to specific linguistic description levels for multilingual and multimodal modeling options of linguistic (linked) data. Furthermore, the activities of this WG provide support for education and educational programs for linguistic data science in a Web-centered context.

In NexusLinguarum linguistic data science is understood as a subfield of the rapidly growing field of data science. Data science can be described as the systematic analysis and study of the structure and properties of data, including methods and techniques to extract knowledge and gain insights from data. The subfield of linguistic data science investigates the analysis, representation, integration, and exploitation of linguistic data for language analysis and language applications. Language analysis spans different linguistic description levels and theoretical bases, e.g. syntax, morphology, terminology, lexicology, etc. Language applications relate to common NLP tasks, e.g. machine translation, speech recognition, sentiment analysis, etc. Linguistic data are typically contained and described in language resources.

Within this context, WG3 aims to cover the broader topic of support for linguistic data science within NexusLinguarum and this document presents first the objectives of WG3, its structure and task leader profiles and the main outcomes and publications that resulted from WG3 activities over the second half of NexusLinguarum from October 2021 to April 2024. Activities of this working group related to the first half of the COST Action NexusLinguarum can be found in Deliverable D3.1 submitted in October 2021. The document then continues to detail the activities for each individual task before providing a final conclusion.

2.1. WG3 Objectives

The title and task of WG3 is to provide support for linguistic data science. The main objective is to offer such support in the form of information of existing and needed resources, approaches and standards. In more detail, the following major objectives have been formulated:

- collect resources and approaches, especially regarding
    - Big Data and LLD
    - Deep Learning and LLD
    - multilingual modeling and LLD
    - multimodal and multidimensional modeling and LLD
- prepare comprehensive state-of-the-art reports on specific topics
- conduct surveys on the utilization and support of deep learning for LLD

- propose, collect, publish and report on modeling of multilingual and multimodal aspects in linguistic data science
- collect and report on skills and competencies important for and in linguistic data science
- propose training programs for linguistic data science

2.2. WG3 Structure and Core Group Profiles

WG3 is one of four WGs within NexusLinguarum with approx. 84 registered members, six task leaders and five tasks. The profiles of WG3 participants range from linguists to computer scientists, including deep learning specialists. An overview of the variety of profiles is reflected in the core group profiles provided below. Unfortunately, Thierry Declerck, our dear colleague and co-WG leader, passed away within the second half of NexusLinguarum and will be sorely missed by all WG3 members and beyond. An overview of the structure of WG3 is provided in Fig. 1.

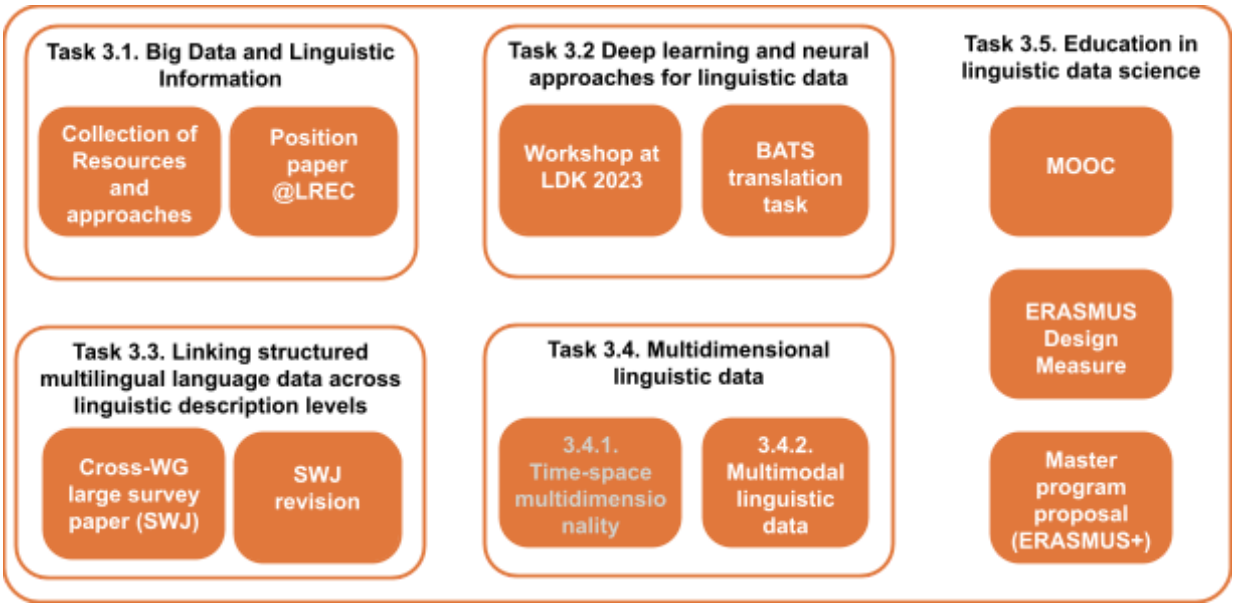

Figure 1: Structure of WG3 and overview of all tasks

The core group of WG3 including their roles, names and countries is presented in Table 1, followed by a detailed description of the core group's profiles.

Role	Person	Country
WG3 leader	Dagmar Gromann	Austria
WG3 co-leader	Thierry Declerck	Germany
Task 3.1 leader	Dimitar Trajanov	North Macedonia
Task 3.2 leader	Radovan Garabík	Slovakia
Task 3.3 leader	Dagmar Gromann	Austria
Task 3.4.2 leader	Ineke Schuurman	Belgium
Task 3.4.2 leader	Thierry Declerck	Germany
Task 3.5 leader	Renato Rocha Souza	Austria
Task 3.5 leader	Rute Costa	Portugal

Table 1: The core group of WG3

Dagmar Gromann, University of Vienna, Austria (WG3 leader and Task 3.3. leader): is Associate Professor at the Centre for Translation Studies (CTS) of the University of Vienna with a focus on computational terminology and language technology. Her research particularly focuses on machine learning and deep learning approaches to multilingual information extraction, including terminological concept systems and cognitive linguistic concepts. She has been project leader of the pilot project "Extracting Terminological Concept Systems from Natural Language Text" (Text2TCS) funded by the European Language Grid (ELG), available as an ELG service, and the project on gender-fair machine translation (GenderFairMT). She was leader of the curricular working group for the development of the new master's program Multilingual Technologies of the CTS in collaboration with FH Campus Wien. In addition, she has organized multiple international scientific events, including as local organizer the 4th Conference on Language, Data and Knowledge (LDK) 2023, which was supported by NexusLinguarum.

Thierry Declerck, DFKI, Germany (Science Communication Manager, WG3 co-leader and Task 3.4.2. co-leader): was senior consultant at the Multilinguality and Language Technology

(MLT) Lab of the German Research Center for Artificial Intelligence (DFKI GmbH) in Saarbrücken, Germany. He was working in a series of European and national projects dealing with a broad range of NLP topics. He was in charge of the DFKI contribution to the H2020 Prêt-à-LLOD Project, which is dealing with the topic of "Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors". Thierry was also instrumental to two W3C Community Groups, Ontology Lexica (Ontolex) and Linked Data for Language Technologies (LD4LT). Together with Dagmar Gromann (and other colleagues), Thierry has co-organized a series of workshops on "Semantic Deep Learning". Thierry was co-chair of the LDK 2021 conference.

Dimitar Trajanov, Ss. Cyril and Methodius University, North Macedonia (Task 3.1. leader): is a Visiting Research Professor at Boston University and a full professor at the Faculty of Computer Science and Engineering. Cyril and Methodius University –Skopje. From March 2011 until September 2015, he was the founding dean of the Faculty of Computer Science and Engineering, and during his tenure, the faculty became the largest technical faculty in Macedonia. After that, from 2015 to 2022, he was Head of the Department of Information Systems and Network Technologies. Dimitar Trajanov is the leader of the Regional Social Innovation Hub, which was established in 2013 as a cooperation between UNDP and the Faculty of Computer Science and Engineering. Dimitar Trajanov is the author of more than 200 journal and conference papers and seven books. He has been involved in more than 70 research and industry projects, of which in more than 40 projects as a project leader. His research interests include Data Science, Machine Learning, NLP, FinTech, Semantic Web, Open Data, Sharing Economy, Social Innovation, Technology for Development, and Climate Change.

Radovan Garabík, Slovak Academy of Sciences, Slovakia (Task 3.2. leader): works in the field of corpus linguistics, natural language processing, large language models and digital lexicography. He was responsible for the design and implementation of written Slovak corpus, corpus of spoken Slovak, Slovak morphology analyser, parallel corpora, Slovak Terminology Database, on-line Slovak language dictionaries portal, and a database of Slovak language linguistic resources. He is also participating on corpus of syntactically annotated Slovak language, corpus of Slovak dialects, several monolingual and bilingual dictionary projects, machine translation projects and is the principal editor and author of several specialized dictionaries and language processing tools and resources. Currently, he represents Slovakia in COST Action CA21167 – Universality, diversity and idiosyncrasy in language technology and the Erasmus+ project "slovake.eu - Extending the e-learning offer with new materials for learning the Slovak language".

Ineke Schuurman, KU Leuven, Belgium (Task 3.4.2 leader): as a voluntary research associate she is member of the Centre for Computational Linguistics (KU Leuven) where she started in 1989 as delegate coordinator of Eurotra-Leuven (MT). Thereafter she was involved in many national, international (Netherlands - Belgium), and European projects. From the beginning (2008) till her retirement she was also involved in CLARIN (ERIC) in several positions. Lately she co-ordinated the EU Competitiveness and Innovation Framework project "Able to Include". Currently she is also involved in the COST Action "advancing Social inclusion through Technology and EmPowerment" (a-STEP) and in the EU project "Sign Language Translation Mobile Application and Open Communications Framework" (SignON).

Rute Costa, Universidade NOVA de Lisboa, Portugal (Task 3.5. leader): is Associate Professor with tenure and habilitation at the Department of Linguistics of the NOVA University Lisbon with a focus on terminology, lexicography and ontologies. Additionally, she is the head of the Linguistic Research Centre (CLUNL) and the research group focusing on Lexicology, Lexicography, and Terminology at NOVA University Lisbon. Engaged in a variety of research endeavours, she has participated and participates in funded projects across several programs including FP7 Program, Horizon 2020, HORIZON Europe, ERASMUS-EDU-2022_EMJM-Design, and the ERASMUS+ Program and the FCT MCTES program, (MORDigital). Notably, she was honoured with the Santander/NOVA Collaborative Research Award 2019-2020, representing NOVA FCSH with the Com@Rehab Project "Communication for interactive rehabilitation in virtual reality." Furthermore, she contributed as a researcher in a Pfizer-funded project titled "Caring Communication: gene therapy in the context of haemophilia." She is scientific coordinator for the PhD Programme in Translation and Terminology and over the course of her career, she has supervised 19 PhD theses and approximately 50 master's dissertations.

2.3. Preliminary Outcomes and Publications

In order to describe our activities in the form of measurable outcomes, this section details events organized within and for WG3, resources collected and provided by WG3, and publications that have resulted from our activities.

Events:

- Workshop on *Linguistic knowledge processing with deep learning (LKPDL 2022)*: This workshop taking place in Jerusalem, Israel on 24 May 2022 was mainly organized within the context of Task 3.2. and took place within the umbrella of the event Nexus Workshops Days in Jerusalem.
- Workshop on *Deep Learning and Neural Approaches for Linguistic Linked Data 2* (*2nd Workshop DL4LD*): This workshop took place in Vilnius, Lithuania in 22 September 2022 within the context of the Nexus yearly meeting and was mainly

organized by Task 3.2.

- *3rd Training School*: Jointly with WG1 this working group organized the 5th Summer Datathon on Linguistic Linked Open Data (https://datathon2023.jezik.hr/) in Lužnica, Croatia from 11. June 2023 to 16. June 2023 with a strong relation to Task 3.2. and a focus on novel neural approaches within the context of linguistic data science. Details on the event are reported in Deliverable D3.2 submitted in November 2023.
- *4th Conference on Language, Data and Knowledge (LDK)*: The leader of WG3 functioned as the main local organizer of this conference taking place in Vienna, Austria from 12. September to 15. September 2023.
- Workshop *Deep Learning, Relation Extraction and Linguistic Data with a Case Study on BATS* (3rd Workshop DL4LD). The workshop took place in Vienna, Austria on 13 September 2023, collocated with the LDK conference. The workshop was mainly organized within Task 3.2 and was focused on the adapting the Bigger Analogy Test Set for other languages.
- Workshop *TermTrends* 2023: co-organized by Rute Costa, task leader of Task 3.5., this workshop took place within the context of the *4th Conference on Language, Data and Knowledge (LDK)* in Vienna, Austria on 13. September 2023.

Future events:

- Workshop on *Deep Learning and Linguistic Linked Data (DLnLD)* represents a continuation of the previous DL4LD workshop activities and will take place in May 2024 collocated with LREC-COLING 2024.
- Workshop *TermTrends* 2024: This workshop will continue its series in June 2024 in Granda Spain.

Resources:

The main resource that was initiated in WG3 and resulted in a NexusLinguarum-wide collaboration across all working groups is a highly multilingual dataset of lexico-semantic relations in 15 natural languages from Bambara, Lithuania and Albanian to Slovak and French. The dataset is called MultiLexBATS, is available on the GitHub of NexusLinguarum and has been published in Gromann et al. (2024b).

Publications:

Here we provide a comprehensive list of publications, which will be described in detail in the respective working groups:

Bączkowska, Anna, and Dagmar Gromann (2023). "From Knobhead to Sex Goddess: Swear Words in English Subtitles, Their Functions and Representation as Linguistic Linked Data." Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje 49, no. 1. DOI: 10.31724/rihjj.49.1.4

Declerck, Thierry, Sam Bigeard, Dorians Callus, Benjamin Matthews, Sussi Olsen, and Loran Ripard Xuereb (2023a). "A uniform RDF-based Representation of the Interlinking of

Wordnets and Sign Language Data." In Proceedings of the 4th Conference on Language, Data and Knowledge, pp. 364-373.

Declerck, Thierry, Sam Bigeard, Fahad Khan, Irene Murtagh, Sussi Olsen, Mike Rosner, Ineke Schuurman, Andon Tchechmedjiev, and Andy Way. (2023b). "A Linked Data Approach for linking and aligning Sign Language and Spoken Language Data." In Proceedings of the Second International Workshop on Automatic Translation for Signed and Spoken Languages, pp. 11-21. European Association for Machine Translation.

Declerck, Thierry, Thomas Troelsgård, and Sussi Olsen (2023c). "Towards an RDF Representation of the Infrastructure consisting in using Wordnets as a conceptual Interlingua between multilingual Sign Language Datasets." In Proceedings of the 12th Global Wordnet.

Declerck, Thierry, and Sussi Olsen (2023). "Linked Open Data compliant Representation of the Interlinking of Nordic Wordnets and Sign Language Data." In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pp. 62-69.

Declerck, Thierry (2022). Integration of Sign Language lexical Data in the OntoLex-Lemon Framework. IDS-Verlag.

Declerck, Thierry (2022). "Towards a new Ontology for Sign Languages." In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 3977-3983.

Declerck, Thierry (2022). Towards the Linking of a Sign Language Ontology with Lexical Data." In Proceedings of Globalex Workshop on Linked Lexicography within the 13th Language Resources and Evaluation Conference, pp. 6-9.

Garabík, Radovan, Ana Ostroški Anić, Sigita Rackevičienė, Giedrė Valūnaitė-Oleškevičienė, Linas Selmistraitis, and Andrius Utka (2023). "Validation of the Bigger Analogy Test Set Translation into Croatian, Lithuanian and Slovak." In Proceedings of the 4th Conference on Language, Data and Knowledge, pp. 402-409.

Garabík, Radovan, ed. (2022) LLOD Approaches for Language Data Research and Management: LLODREAM2022: International Scientific Interdisciplinary Conference, September 21-22, 2022: Abstract Book. Mykolas Romeris University. https://llodapproaches2022.mruni.eu/wp-content/uploads/2022/10/LLOD_2022-Book-of-Abstracts.pdf

Gromann, Dagmar, Elena-Simona Apostol, Christian Chiarcos, Marco Cremaschi, Jorge Gracia, Katerina Gkirtzou, Chaya Liebeskind, Liudmila Mockiene, Michael Rosner, Ineke Schuurman, Gilles Sérasset, Purificação Silvano, Blerina Spahiu, Ciprian-Octavian Truică, Andrius Utka, and Giedre Valunaite Oleskeviciene (2024a). Multilinguality and LLOD: A survey across linguistic description levels. Semantic Web Journal, IOS Press. DOI: 10.3233/SW-243591

Gromann, Dagmar, Hugo Gonçalo Oliveira, Lucia Pitarch, Elena-Simona Apostol, Jordi Bernad, Eliot Bytyçi, Chiara Cantone, Francesca Frontini, Radovan Garabik, Jorge Gracia, Litizia

Granata, Anas Fahad Khan, Timoteij Knez, Penny Labropoulou, Chaya Liebeskind, Maria di Buono, Ana Ostroški Anić, Sigita Rackevičienė, Ricardo Rodrigues, Gilles Sérasset, Linas Selmistraitis, Sidibé Mahammadou, Purificação Silvano, Blerina Spahiu, Enriketa Sogutlu, Ranka Stanković, Ciprian-Octavian Truică, Giedrė Valūnaitė Oleškevičienė, Slavko Zitnik, and Katerina Zdravkova (2024b). "MultiLexBATS: Multilingual Dataset of Lexical Semantic Relations." In Proceedings of LREC-COLING 2024. https://zenodo.org/records/10956565

Mishev, Kostadin, Ana Gjorgjevikj, Irena Vodenska, Lubomir T. Chitkushev, and Dimitar Trajanov (2020). "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers." IEEE access 8: 131662-131682.

Trajanov, Dimitar, Vangel Trajkovski, Makedonka Dimitrieva, Jovana Dobreva, Milos Jovanovik, Matej Klemen, Aleš Žagar, and Marko Robnik-Šikonja. (2023). "Review of Natural Language Processing in Pharmacology." *Pharmacological Reviews* 75, no. 4: 714-738.

Trajanov, Dimitar, Elena-Simona Apostol, Radovan Garabík, Katerina Gkirtzou, Dagmar Gromann, Chaya Liebeskind, Cosimo Palma, Michael Rosner, Alexia Sampri, Gilles Sérasset, Blerina Spahiu, Ciprian-Octavian Truică, and Giedre Valunaite Oleskeviciene (2024). "From Linguistic Linked Data to Big Data." In Proceedings of LREC-COLING 2024. https://zenodo.org/records/10956967

Rackevičienė, Sigita, Liudmila Mockienė, Andrius Utka, and Aivaras Rokas (2021). "Methodological Framework for the Development of an English-Lithuanian Cybersecurity Termbase." *Studies about Languages* 39, 85-92.

Rackevičienė, Sigita, Andrius Utka, Agnė Bielinskienė, and Aivaras Rokas (2022). "Distribution of Terms Across Genres in the Annotated Lithuanian Cybersecurity Corpus." *Respectus Philologicus* 41, no. 46, 26-42.

Rosner, Michael, Sina Ahmadi, Elena-Simona Apostol, Julia Bosque-Gil, Christian Chiarcos, Milan Dojchinovski, Katerina Gkirtzou et al. (2022). "Cross-Lingual Link Discovery for Under-Resourced Languages." In 13th International Conference on Language Resources and Evaluation (LREC), JUN 20-25, 2022, Marseille, France, pp. 181-192. European Language Resources Association ASSOC-ELRA.

Rizinski, Maryan, Hristijan Peshov, Kostadin Mishev, Milos Jovanovik, and Dimitar Trajanov (2024). "Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (XLex)." IEEE Access .

Schuurman, Ineke, Thierry Declerck, Caro Brosens, Margot Janssens, Vincent Vandeghinste, and Bram Vanroy (2023). "Are there just WordNets or also SignNets?." In Proceedings of the 12th Global Wordnet Conference, pp. 172-178.

2.4. Collaborative and Joint Activities with other Groups and Initiatives

WG4 & WG1

Within the context of Task 3.1. on Big Data and Linguistic Linked Open Data, we closely collaborated with members of WG4 and also WG1 in order to identify several use cases for this mutually beneficial union of two important technologies. The results of this collaboration are published in Trajanov et al. (2024). With WG1 we closely collaborated on the evaluation of cross-lingual linking methods and approaches published in Rosner et al. (2022).

Cross-WGs

There are four main initiatives that truly united members from all working groups: (1) the creation of a multilingual dataset of lexico-semantic relations called MultiLexBATS started from Task 3.2, (2) a survey on the current status of linguistic description levels in LLOD organized within Task 3.3, (3) a survey on the current status of neural approaches in connection with LLOD started from WG2 and strongly tied to Task 3.2 and other WG3 topics and members, and (4) educational activities within Task 3.5. The first two and the fourth initiatives will be described in detail in the descriptions of the individual tasks activities, while the third initiative will be detailed in the deliverable of WG2. The educational activities comprised the organization of a Massive Open Online Course on linguistic linked data, the development of a full curriculum for a master's program on the same topic reported as D3.3 in April 2024, as well as the submission of a ERASMUS+ master's program proposal.

OntoLex-Frac

WG3 closely collaborates with members of the W3C Community Group OntoLex, especially with members of the interest group on OntoLex module for FRequency, Attestations and Corpus data (FRAC) on the topic of multimodal modeling of linguistic linked data (Task 3.4.2 of WG3). Furthermore, additional collaborations with OntoLex members on modeling time and space in reference linguistic linked data have been initiated.

OntoLex-Morph

WG3 closely collaborates with members of the W3C Community Group OntoLex-Morph in connection with Task 3.3. In particular, Elena-Simona Apostol, Katerina Gkirtzou, and Ciprian-Octavian Truica actively contributed to the development of a morphology module for the already existing OntoLex module to enhance the possibility to represent morphological perspectives and data.

Thierry Declerck was invited on the 6th of October 2021 to give a talk entitled “About the Linguistic Linked Open Data initiative” at the plenary meeting of the Multi3Generation COST Action CA18231 (https://multi3generation.eu/). The purpose of this invitation was to investigate potential lines of cooperation between Multi3Generation and NexusLinguarum. Two topics for possible cooperation were immediately recognized: the encoding of Sign Languages and the use of knowledge graphs for the generation of common sense inferences.

Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letiția Pârcălăbescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, Elena Lloret, **Elena-Simona Apostol**, **Ciprian-Octavian Truică**, Branislava Šandrih, Albert Gatt, Sanda Martinčić-Ipšic, Gábor Berend, Gražina Korvel. *Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning*, Journal of Artificial Intelligence Research, 73:1131-1207, April 2022. DOI: 10.1613/jair.1.12918

LD4LT meetings

NexusLinguarum members are involved in and leading a W3C Community Group: The Linked Data for Language Technologies (LD4LT) that was founded in the previous FP7 project “LIDER'' and that is used for the broader discussion of issues related to linked data and its applications in NLP. NexusLinguarum partners are participating to the regular LD4LT telcos.

BPMLOD

Within the W3C community group on Best Practices for Multilingual Linked Open Data Community Group (BPMLOD), there are several initiatives and subgroups to which WG3 members and results contribute. The working group on Neuro-Symbolic LLOD connects to Task 3.2. and several members of WG3 actively contribute, also to Licensing and Metadata as well as Sign Language strongly tied to Task 3.4.2 and emerging from there.

EASIER, SignON

Within the context of Task 3.4.2 there is vivid exchange with sign language projects on the topic of modeling and exchanging sign language data with LLOD. In particular, members of the projects EASIER (EU Horizon 2020 grant number 101016982) and SignON (EU Horizon 2020 grant number 101017255) presented their works in Task 3.4.2 meetings and contributed their expertise to the discussions and work of this task. Results of this vivid collaboration can be seen in publications such as Declerck et al. (2023b).

3. WG3 Task Activities

This section provides details on all WG3 activities structured by task, providing information on the task leader, a brief overview of the task, and then detailed explanations of all activities.

3.1. Task 3.1. Big Data and Linguistic Information

Task leader: Dimitar Trajanov

Overview:
In this task, Big Data sources and state-of-the-art statistical analysis are studied in combination with LLOD in order to better understand language. Visual analytics will be also considered for this task. This will have an impact on all subdomains of linguistics, from typology to syntax to comparative linguistics.

Activities:
The initial activities of this task were centered on collecting a complete overview of existing resources, approaches, and publications on the union of Big Data and Linked Data. While a large number of publications on the topic of Big Data and Knowledge Graphs could be found, e.g. Janev et al. (2020a, 2020b), very few approaches addressed Linked Data in connection with Big Data and none could be identified on the specific union with Linguistic Linked Data. Thus, instead of providing a survey or review, the leader and members of this task decided to publish a position paper (Trajanov et al. 2024) on the benefits of this union and a detailed description of use cases, from enhancing existing large resources like DBnary and corpora access to information extraction.

3.2. Task 3.2. Deep learning and neural approaches for linguistic data

Task leader: Radovan Garabík

Overview:
Currently, deep learning techniques have gained popularity in many research areas, NLP being one of them. In fact, the field is being revolutionized by the emergence of relatively available huge language models based on attention (transformers), originally BERT, with the focus shifting towards GPT-based models later on. The goal of this task was to study the effective use of deep learning in understanding the specifics of linguistic data in a big data context, to be better exploited and combined with linked data mechanisms.

Within the task, we organized several workshops focused on different topics:

The workshop “Linguistic knowledge processing with deep learning” was a hybrid workshop (physical & online) held in Jerusalem, Israel on 24 May 2022, as part of the Nexus Workshops days in Jerusalem.
The topics for the workshop were:

- Linguistic Knowledge Extraction / Discovery
- Machine Translation
- Multilingual linguistic data processing
- Word Sense Disambiguation and Entity Linking in LLOD
- Terminology and Knowledge Management

The workshop included two presentations and a BATS Translation Task hands-on session.
The address of the workshop web page is
https://www.jct.ac.il/כנסים/nexus-workshops-days-in-jerusalem/linguistic-knowledge-processing-with-deep-learning

The second “Workshop on Deep Learning and Neural Approaches for Linguistic Linked Data 2 (2nd Workshop in the DL4LD series)” was held in conjunction with the conference “LLOD approaches for language data research and management” (LLODREAM2022) in Vilnius, Lithuania on 22 September 2022 also as a hybrid workshop. The workshop included three presentations. The webpage of the workshop is https://llodapproaches2022.mruni.eu/?page_id=239 and the workshop abstracts are published in the LLODREAM2022 Book of Abstracts:
https://llodapproaches2022.mruni.eu/wp-content/uploads/2022/10/LLOD_2022-Book-of-Abstracts.pdf

The workshop “Deep Learning, Relation Extraction and Linguistic Data with a Case Study on BATS” (the 3rd Workshop in the DL4LD series) was held in conjunction with the LDK2023 conference in Vienna, Austria on 13 September 2023. The workshop included 5 presentations and one invited talk. The webpage is http://dl4ld2023.mruni.eu/ and the articles are published in the conference proceedings available at
https://aclanthology.org/volumes/2023.ldk-1/

Related activity was the “5th Summer Datathon on Linguistic Linked Open Data”, organized jointly between WG3 and WG1 in Lužnica, Croatia from 11. June 2023 to 16. June 2023. The focus of the datathon was on novel neural approaches within the context of linguistic data science. The webpage of the datathon is https://datathon2023.jezik.hr/.

Furthermore, originating from this task, we started a cross-NexusLinguarum activity of generating a multilingual dataset of lexico-semantic relations. The dataset called MultiLexBATS is available on the GitHub page of NexusLinguarum. The idea behind the dataset was to probe and evaluate the possibility to acquire relations in multiple languages from neural language models. The results of these experiments with neural language models on MultiLexBATS are presented in the publication Gromann et al. (2024b). The dataset is currently available in 16 languages from the Indo-European, Afro-Asiatic and Mande language families: Albanian, Greek, German, French, Italian, Portuguese, Spanish, Romanian, Lithuanian, Slovak, Croatian, Macedonian, Serbian, Slovenian, Hebrew, and Bambara. The dataset consists of the following different types of lexico-semantic relations: hypernymy, hyponymy, meronymy, synonymy, and antonymy. We tested the ability of neural language models to predict the right target word in a relation triple across all of these languages with mBERT, XLM-R and BLOOM as a generative pretrained transformer. The results showed models showed better performances on languages explicitly included in the training settings. Furthermore, we proposed a new task of utilizing the idea of analogies, such as *king is to queen as man is to woman*, to predict the translations, e.g. *apple is to fruit as manzana es como…* with the expected result to be *fruta*. This task turned out to be rather difficult with a low proportion of correct predictions, however, more experiments and investigations with translation-related tasks on this dataset would be a desideratum. In addition, we intend to investigate lexical gaps that become evident on the basis that this dataset is aligned via the English language in order to permit cross-lingual experiments, such as the analogy-based translation task described above.

3.3. Task 3.3. Linking structured multilingual language data across linguistic description levels

Task leader: Dagmar Gromann

Overview:
This task focuses on how data for the basic levels of phonology, morphology and lexicon, often spread across datasets of varying extent, quality and format, can be described, stored and accessed uniformly.

Activities:
Starting from the idea of providing a best practice for modeling different linguistic description levels in relation to LLD, it quickly became clear there is no recent and comprehensive survey to provide a solid basis for a best practice. Thus, members of all working groups came together to provide a very comprehensive and state-of-the-art review

on multilinguality and LLOD with a particular focus on different linguistic description levels published at the Semantic Web Journal (Gromann et al. 2024a).

Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, a total of 16 experts jointly analyzed and evaluated an initial set of 25,074 publications quickly reduced to 210 for manual screening. We then annotated the top-most 210 papers regarding their relevance as well as potential subtopic of this topic. Thereby, we could identify the best ranked and humanly determined top 112 papers that were then clustered by subtopic in order to start a detailed reading and systematic review. To further validate this automated search, we also compiled a repository of publications that these experts considered central to this topic that was compared against the automatically generated results with a very decent overlap, that is, 80% of the publications considered relevant by human experts could also be retrieved by our automated search and semi-automated ranking method.

We decided to determine the number of experts needed for reviewing and summarizing the identified subtopics based on the number of papers that resulted in each, which are represented in Table 1. For the largest topic, OntoLex, three experts were assigned, while for smaller subtopics one to two experts reviewed the resulting papers.

Subtopic designation	No Papers	No Experts
application	15	2
BabelNet	5	1
Literature review	5	1
LLOD infrastructure	4	1
morphology	5	1
ontolex-lemon	25	3
overview paper	6	1
representation	12	2
resources	12	2
standards	5	1
under resourced languages	4	1
use cases	12	2

Table 2: Types of subtopics and number of papers of the result set as well as experts assigned to each subtopic

The final summaries for each linguistic description level and cluster was further complemented with an additional search of potentially relevant publications by the annotation and expert groups. We identified the main linguistic description levels to this date covered by existing approaches, which are represented in Fig. 5 where the size of the bubble approximates its coverage in existing publications.

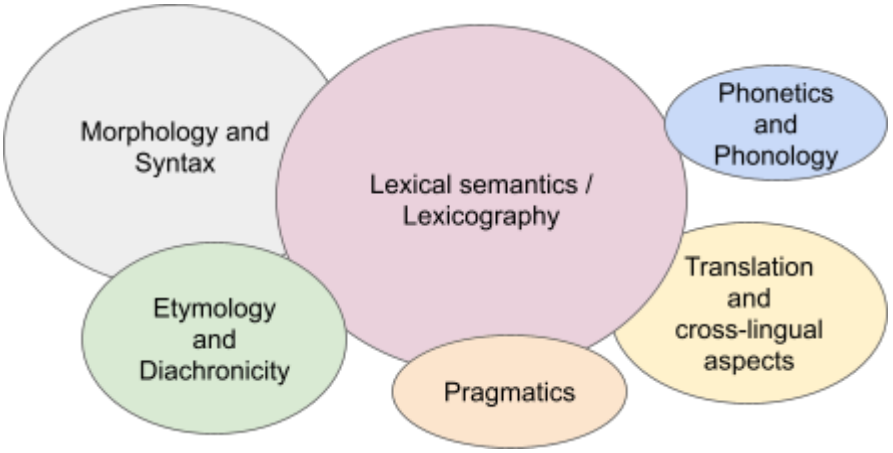

Figure 5: Representation of linguistic description levels in final result set

As can be seen from Fig. 5, the linguistic description levels that are very well covered within LLOD in a multilingual context are lexical semantics, morphology and syntax, and some fewer approaches exist for translation and cross-lingual approaches. For the two levels phonetics and phonology and etymology and diachronicity fewer publications and approaches could be identified. The level that requires most future research is pragmatics, where only very few publications on discourse markers in a multilingual LLOD context could be found.

3.4. Task 3.4. Multidimensional linguistic data

The original idea of this task as proposed initially was to focus on various dimensions of linking language resources, e.g. time and sociolect, in order to foster diachronic and sociolinguistic interoperable research across different sources in LLOD. However, it turned out to be practically very difficult to define these multiple dimensions within one task and some dimensions, especially diachronic aspects, strongly overlapped with use cases in WG4. Thus, we finally decided to split Task 3.4. into two tasks that foster the clarity of activities and reduce overlap with other activities in NexusLinguarum. The first subtask will provide a container task to join activities of WG1 and WG4 on time and space. The second subtask focuses on multimodal modeling of LLD.

3.4.1. Task 3.4.1. Time-space multidimensional linguistic data

This subtask has been covered by WG1 and WG4 use cases, which is why no further initiatives on this topic were started within WG3. We thus refer to the activity reports of these two working groups for details on this topic.

3.4.2. Task 3.4.2. Multimodal linguistic data

Task leaders: Ineke Schuurman, Thierry Declerck

Overview:
One central aspect of multidimensionality within the context of linguistic data science is being able to represent, interlink, and exchange multimodal linguistic data. Multimodal here refers to the existence of more than one modality, where modality is defined as audio, visual, textual, or other channel. In the beginning several topics were discussed, but the field was so broad that we were not able to do justice to all of them. After some time (early 2022) it was decided to pay special attention to Sign Languages.

Activities:

In July 2022, several people not involved in NexusLinguarum gave short presentations on their work relating/linking WordNet and Sign Languages, amongst them people involved in the EU projects EASIER (Sam Bigeard and Anna Vacalopoulou) and SignON (Ineke Schuurman, also involved in NexusLinguarum). On April 19th 2023, Irene Murtagh (TU Dublin, involved in SignON) gave a presentation on Sign_A, the notation system for SLs she developed (with focus on Irish Sign Language).

During monthly meetings interested NexusLinguarum participants, mostly around 8 persons, discussed the issues we are to face when dealing with Sign Languages. One of them being that there is no generally accepted writing system, not one used by Deaf Communities, neither by institutions, researchers, … in the field. A second issue is that Sign Languages are very under-resourced languages, even for example those used in the UK and the US (two really different ones!). Often glosses are used, a kind of representative label written in capitals assigned to a sign. Often, a notation system is not being used. As outsiders (not being (near-)native speakers of an SL) we are to accept this. But even when the notation system is being used, glosses together with 'suggested translations' in a surrounding spoken language are of importance, for example when linking with WordNets, There are contacts with the people behind for example Open Multilingual WordNet.

Thierry Declerck had taken several initiatives concerning Sign Languages, especially also in other projects and collaborations. After he suddenly passed away we had to trace his work, not always with success. We are now in a smaller group working on Sign Languages, and will continue doing so after NexusLinguarum finished (BPMLOD, with focus on Sign Languages

and WordNets). Currently we are working on an overview of the state of affairs concerning SLs and their use in the digital age.

3.5. Task 3.5. Education in linguistic data science

Task leaders: Rute Costa

Overview:

To introduce linguistic data science in a cross-disciplinary academic infrastructure, in Task 3.5 we developed four main activities: (1) Assessment of the existing courses and educational programs related to linguistics and data science; (2) Designing and implementing a Massive Open Online Course (MOOC) on Linguistic Data Science; (3) Linking Linguistics to Data Science (LL2DS) - ERASMUS-EDU-2022-EMJM-DESIGN and (4) Erasmus Mundus Joint Masters in Linguistic Data Science (EMLDS).

Assessment of the existing courses and educational programs related to linguistics and data science

To foster the integration of linguistic data science into a cross-disciplinary academic framework, NexusLinguarum undertook the development of a curriculum for a Europe-wide master's degree program. Despite the expansive nature of the field, characterized by diverse perspectives and educational methods, an assessment of existing curricula was conducted to identify shared skills and competencies. Subsequent steps will entail crafting the course structure as part of an Erasmus+ initiative.

As an initial and ongoing activity in Task 3.5. we continuously collected existing courses and educational programs related to linguistics, data science, and especially linguistic data science as exemplified in Table 2. For a detailed analysis of existing courses and programs, we collected their main skills and competencies, such as related to linguistics, programming, software design, language technologies, and information management.

#	Source / Institution	Course	Level
1	Universität Bern	Introduction in Digital Humanities - Text Digital–Von der Aufbereitung zur Auswertung	BA
2	Austrian Centre for Digital Humanities and Cultural Heritage	ACDH Tool Galleries	-
3	International University of La Rioja	Máster Universitario en Humanidades Digitales	MA
4	Université de Strasbourg	Master Technologies des Langues	MA
5	University of Macerata	PhD in Humanities and Technologies	PhD
6	European Consortium	European Masters Program in Language and Communication Technologies	MA
7	Univ Gothenburg	Master Language Technology	MA
8	Univ. Oslo	Master Language Technology	MA
9	MLT Carnegie Mellon Univ	Master Language Technology	MA
10	FTI Univ. Geneva	Master in Multilingual Communication Technology	MA
11	University of Helsinki	Master in Linguistic Diversity and Digital Humanities	MA
12	Stanford Data Science Initiative (SDSI)	data science for linguistics	Extension
13	MOOC	Introduction to a Web of Linked Data	Extension
14	MOOC - Coursera	Web of Data	Extension
15	University of Helsinki	Digital Humanities and Social Sciences	MA
16	University of Vienna and Applied University Campus Vienna	Multilingual Technologies	MA
17	Universidade NOVA de Lisboa	Terminology and Management for Special Purposes (Linguistics)	MA
18	Universidade NOVA de Lisboa	Linguistics - specialization in Lexicography and Terminology	PhD
19	University of Helsinki	Master in Contemporary Societies	MA
20	University of Helsinki	Introduction to Digital Humanities and Social Sciences	BA
21	University of Luxembourg	Introduction to Computational Text Analysis and Text Interpretation	BA
22	University of Luxembourg	Computing Culture. An introduction to Python programming for the humanities	PhD
23	University of Luxembourg	Introduction to Digital History	BA, MA
24	University of Ljubljana, University of Zagreb	Digital Linguistics	MA
25	University of Bucharest	Digital Humanities	MA
26	University of Belgrade	Digital Humanities	MA
27	University of Porto-Faculty of Arts and Humanities	Studies in Language Sciences-Human Language Technologies	PhD
28	University of Porto-Faculty of Sciences	Data science	MA
29	University of Gdansk, Poland	English Philology-Natural Language Processing	MA
30	Vytautas Magnus university, Lithuania	Digital Humanities	BA

Table 3: Institutions with educational initiatives related to linguistic data science

Some topics, such as Semantic Web Technologies, Data Mining, and Information Extraction, recurred more frequently than other topics across courses analyzed.

The topics of the list in Fig. 6 were then clustered into thematic groups (e.g. Semantic Web; NLP Pipelines; Coding, Machine Learning & Databases; etc.) and analyzed by frequency of appearance. Some topics would occur in more than 75% of the programs (e.g. “Semantic Web Technologies, Linked Data, SPARQL”; “Text/Data Mining & Information Extraction”) while others would appear less frequently (e.g. “Handwritten Text Recognition (HTR)”, “Statistics”), according to the program specificities, overarching department/institution and level (e.g. MA, BA, PhD). Some programs have prerequisites, which would demand a deeper analysis on necessary skills. A companion analysis of the software platforms, frameworks and programming languages was also made, yielding the following list of tools: (Oxygen, Tesseract, Transkribus, Git, Protégé, XMLMind, Gephi, R, Sonic, Visualiser, Wordnet, Framenet, Lexonomy, Sklearn - Python package).

Massive Open Online Course (MOOC) on Linguistic Data Science

To spread the accumulated expertise resulting from the collaborations and joint initiatives within NexusLinguarum to a wider audience, the organization of a Massive Open Online

Course (MOOC) on the topic of linguistic data science was initiated. An overview of the schedule and topics as well as the lecturers of the MOOC can be found in Table 4.

No.	Module	Lecturer
1	Welcome and Introduction	Jorge Gracia
	Installing: VocBench, Jena, Protégé, etc.	Max Ionov
	Semantic Web and Linked Data in a nutshell	Jorge Gracia
2	Linguistic Linked Data	Christian Chiarcos
	Modeling: Ontolex lemon (+ turtle examples)	John McCrae
	Tools for ontolex lexicon building: VocBench	Armando Stellato
3	LD tools (Jena, Protégé, …)	Max Ionov
	SPARQL	Slavko Žitnik
4	LD generation (RML, VocBench)	Andon Tchechmedjiev Armando Stellato
	Corpora and annotation (NIF, Web Annotation, …)	Christian Chiarcos
5	Metadata	Penny Labropoulou Andon Tchechmedjiev
	Resources: DBnary, Wikidata, ...	Gilles Sérasset
	LD and lexicography	Jorge Gracia
6	LD and terminology	Sara Carvalho
	Deep learning and LLD	Dagmar Gromann
7	Use case: LiLa	Marco Passarotti Francesco Mambrini

Table 4: Schedule and Lesson overview of MOOC on Linguistic Data Science

The WG3 member Slavko Zitnik took over the task of organizing the initiative. All materials including slides, trial recordings, quizzes, assignments, and screencasts were centrally collected in NexusLinguarum repositories by December 2023 and had to be finalized by February 2024. Furthermore, all lecturers produced an exact transcript for their lectures to be prepared and time-efficient for the professional video recording.

The professional video recording was collocated with the last MC meeting in Athens in March 2024 and organized by the local organizers. The recordings of the lectures were organized in time slots over a total of two days. While the first version of the videos is already available, the videos, lecturer positioning on or next to the slides and the presentation of quizzes and assignments are currently being finalized to be completed by the end of NexusLinguarum.

Linking Linguistics to Data Science (LL2DS) - ERASMUS-EDU-2022-EMJM-DESIGN.

Linking Linguistics to Data Science (LL2DS) project submitted in February 2022, and approved in July 2022, aimed at designing a new study programme at Master level in the field of linguistic data science (LDS). LL2DS was launched on the 1st of October 2022 and its funding period finished on the 31st of December 2023. As a mono-beneficiary project, NOVA University Lisbon NOVA University Lisbon is the coordinator of the project.

This project arose from the collaboration generated within the COST Action CA18209 - European Network for Web-centred Linguistic Data Science (NexusLinguarum). As a reminder, the main goal of this COST Action, as stated in its Memorandum of Understanding (MoU), is to "promote synergies across Europe between linguists, computer scientists, terminologists, language professionals, and other stakeholders in industry and society, to investigate and extend the area of linguistic data science." Building on complementary initiatives such as the H2020 ELEXIS and Prêt-à-LLOD projects, as well as the ERC-awarded LiLa project, NexusLinguarum has been working on modeling, interlinking, evaluating, and enriching multilingual language resources to make them available to academia, industry, other institutions, and citizens in general, in a machine-processable way, thus contributing towards the development of an ecosystem where knowledge can be linked and shared, in open access, across languages and domains. To address the ongoing challenges of semantic interoperability while preserving - and fostering - linguistic diversity, this growing linguistic data science (LDS) community has been focusing on the development - and enhancement - of standards and best practices, in line with the FAIR principles. Given the multidisciplinary nature of LDS, and the pivotal role of linguistic resources towards a more empowered, active, and digitally literate society, it is believed that the profiles emerging from the new Master's study programme would be highly sought after, both in academia and outside of it. Furthermore, LL2DS would contribute to stabilize and internationalize *curricula* and teaching practises in an innovative research area, with a promising and flourishing community.

The general objective of LL2DS was to have, at the end of the project, the design of a fully integrated Erasmus Mundus Joint Master Degree (EMJMD), delivered by a consortium involving the NOVA University of Lisbon (Portugal) the Università Cattolica del Sacro Cuore of Milan (Italy), the University of Zaragoza (Spain), the University of Ireland (Galway) and the National Institute for Research in Digital Science and Technology – INRIA (France). LL2DS's goal was, thus, to position itself as innovative amongst programmes around the world in its novel approach of linking Linguistics to Data Science in terms of its methodology, intellectual content, pedagogical approach, and curricular structure. Given the wide range of competencies in the area of LDS, the fundamental approach of the LL2DS project was to exploit the specific expertise in research and teaching in linguistics, data science and computer science provided by all the partners involved. This will also increase the

competitiveness of linguistics, data science and computer science as a discipline by firmly linking it to the shifting demands of the academic and professional job market.

We also undertook a Competitive Audit and Needs Survey and asked the company we hired to develop a market study to lay the groundwork for the creation of the proposed course. To achieve this and consolidate the curriculum design, successfully to both student needs and market demands, this study synthesizes the main findings: (1) benchmarking analysis regarding the existing educational offer by European and North American institutions in the field of linguistic data science; (2) the results of the surveys conducted, to which academics, potential students, and potential employers, provided valuable insights regarding the interest and opportunities of such program; (3) maps the career paths of graduates from the master's program in linguistic data science and identifies potential employers in the field; and (4) the systematisation of the overall conclusions of the study.

4. Erasmus Mundus Joint Masters in Linguistic Data Science (EMLDS)

As a result of the LL2DS, in February 2024, we submitted a fully designed ERAMUS MUNDUS PROGRAM, entitled Linguistic Data Science (LDS). There are four full partners involved in the EMLDS: NOVA University Lisbon (Portugal) (NOVA) as coordinator, University of Zaragoza (Spain) (UNIZAR), Università Cattolica del Sacro Cuore (Italy) (UNICATT) and University of Galway (Ireland) (UoG) as full partners. The involvement of associated partners is also essential for the consortium. These are INRIA – National Institute for Research in Digital Science and Technology (France), University of Aveiro (Portugal) and Lexicala by K Dictionaries (Israel). The Consortium also established the role of External Partners. These are not part of the Consortium but will take part in a. disseminating the programme, b. facilitating knowledge and skills transfer, and c. supporting possibilities for secondment or internship placement. An initial list of External Partners was proposed by each full partner and agreed upon during Consortium meetings. Their support is reflected in the External Partners letters of support under the Other Annexes section. These partners are the Faculty of Mathematics and Physics at Charles University (Czech Republic), Semantic Web Company (Austria), Datrix AI Solutions Group (Italy), Universidad Politécnica de Madrid (Spain), Priberam Informática (Portugal), and Unbabel Inc (USA).

Linguistic data science is considered by the community as a subfield of data science, with linguistics playing a key theoretical and methodological role in the extraction of both structured and unstructured knowledge from language resources, thereby facilitating its organization, representation, and sharing at different levels of analysis (e.g. morphological, lexical, semantic, syntactic).

In addition to linguistics, specific related areas such as computational linguistics, corpus linguistics, data modeling, natural language processing (NLP), artificial intelligence, and deep learning, are also critical in linguistic data science, supporting the in-depth processing and encoding of considerable amounts of linguistic data while ensuring its (re)usability in multilingual applications, as well as in different domains (e.g., healthcare, digital humanities, lexicography, finance).

To further strengthen this community's contribution to the linguistic digital landscape, especially - though not exclusively - in Europe, it is paramount to provide formal education programmes within the scope of the Linguistic Linked Open Data (LLOD) paradigm, aimed not only at new generations of researchers but also at people who wish to retrain in complementary or other, related fields. Given the multidisciplinary nature of linguistic data science, and the pivotal role of languages and language resources towards a more empowered, active, and digitally literate society, we believe belief that the profiles emerging from the EMLDS will be highly sought after, both in academia and outside of it.

Furthermore, the master programme would contribute to stabilize and internationalizing curricula and teaching practices in an innovative research area, with a promising - and flourishing - community.

EMLDS is a 120 ECTS English-language programme awarding a Joint Degree. Based on the collaboration amongst four European universities and 3 associated partners, the EMLDS master aims at providing a novel approach to linking linguistics to computer sciences and data science in terms of its methodology, scientific content, pedagogical approach, and curricular structure. The EMLDS has three different objectives:

1) to increase the competitiveness of linguistics, computer science and data science as disciplines, by firmly linking them to the shifting demands of the academic and professional job market, specifically the growing need for accurate, accessible, and inspiring knowledge;

2) to strengthen linguistic data science as an educational and research field, given its cross-disciplinary framework and cutting-edge nature, aiming to respond to contemporary social and technological challenges;

3) to equip students with the methodological and critical instruments to assess fundamental technical, ethical and political implications of linguistic data science at the local, national, regional, European and global levels.

The successful completion of the EMLDS leads to the award of a joint degree, duly accredited by the countries where the degree-awarding higher education institutions are based. It brings together the expertise and experience of the following master programmes:

The Master in Language Sciences - Terminology and Computer Lexicography by NOVA; the Master in Informatics Engineering by UNIZAR, the Master in Linguistic Computing at UNICATT, and the Master in Data Analytics at UoG. In terms of transdisciplinarity, the EMLDS master combines the curricula of four programmes in this field, with the study plan being delivered by different academic departments namely: Department of Linguistics at NOVA FCSH, Department of Linguistic Sciences and Foreign Literature at UNICATT, the Department of Computer Science at UNIZAR and Data Science Institute at UoG.

The EMLDS two-year programme is divided into four components: (1) Induction Days, (2) coursework in Semester 1 and Semester 2, (3) thematic training sessions and orientation seminar in Semester 3, to be held on a rotating basis by each of the Consortium members, and (4) master dissertation, internship with report or project in Semester 4, in the university of the student's choice. Students will also have the possibility of attending Portuguese, Spanish, Irish and Italian courses during the semester they are at NOVA, UNIZAR, UoG or UNICATT, respectively. Courses and semester distribution Students can follow four different mobility paths, as described ahead in the "curriculum design" section. Each path has its own curricular specificities and its own balance of the involved scientific areas, so the student can choose the path that accommodates best to their background and interests. Mobility paths are proposed by each candidate at application stage and are afterwards confirmed by the Joint Admissions Committee.

Figure 1: Mobility paths

All 46 curricular units have been submitted. The submitted forms contain among other information, the (1) learning outcomes of the curricular unit and their compatibility with the teaching methods (knowledge, skills and competencies to be developed by students), (2) syllabus, (3) demonstration of the syllabus coherence with the curricular unit's objectives, (4) specific teaching and learning methodologies of the curricular unit articulated with the model, (5) evaluation, (6) demonstration of the coherence between the teaching methodologies, and (7) learning outcomes and the bibliography. Each curricular unit has been assigned to at least one teacher.

We will be informed in July if EMLDS has been awarded.

4. Conclusion and Prospects

This report details all major activities accomplished in WG3 within the last two and a half years from October 2021 to April 2024, major collaborations established, major outcomes of these activities, and planned future activities reported in each task description. Apart from all these formal and measurable activities, we would also like to mention the networking and community building effect of this COST Action. In addition to large- and small-scale partnerships on conducting very specific work, jointly preparing publications, or organizing events, we have also established a very friendly and supportive network on the topic of support for linguistic data science, in which expertise is as much exchanged as friendly, amicable conversation (online, offline in meetings and events, and per e-mail).

While the regular meetings will discontinue with the end of NexusLInguarum, several events have established themselves within the course of the action and will be continued, which includes especially future Language, Data and Knowledge (LDK) conferences, the next being planned for 2025, workshops such as DLnLD established in Task 3.2. and TermTrends, which already continue in 2024, the 3rd training school and 5th datathon on LLOD is planned to be continued. Apart from these scientific and collaborative events, there are ongoing video conferences on sign languages, multimodality, neuro-symbolic LLOD and licensing among other topics are being continued already within the context of the W3C Best Practices for Multilingual Linked Open Data Community Group (BPMLOD). This is particularly useful as a platform in which to discuss and contribute NexusLinguarum results in form of best practices. Given the strong collaborations and established interdisciplinary community ties across working groups and NexusLinguarum members, many initiatives continue with regular online meetings with the goal to enhance started approaches, continue initiated work building on NexusLinguarum, and work on challenges and gaps identified in the multiple comprehensive review and survey papers.

References

WG 3 publications

Bączkowska, Anna, and Dagmar Gromann (2023). "From Knobhead to Sex Goddess: Swear Words in English Subtitles, Their Functions and Representation as Linguistic Linked Data." Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje 49, no. 1. DOI: 10.31724/rihjj.49.1.4

Declerck, Thierry, Sam Bigeard, Dorians Callus, Benjamin Matthews, Sussi Olsen, and Loran Ripard Xuereb (2023a). "A uniform RDF-based Representation of the Interlinking of Wordnets and Sign Language Data." In Proceedings of the 4th Conference on Language, Data and Knowledge, pp. 364-373.

Declerck, Thierry, Sam Bigeard, Fahad Khan, Irene Murtagh, Sussi Olsen, Mike Rosner, Ineke Schuurman, Andon Tchechmedjiev, and Andy Way. (2023b). "A Linked Data Approach for linking and aligning Sign Language and Spoken Language Data." In Proceedings of the Second International Workshop on Automatic Translation for Signed and Spoken Languages, pp. 11-21. European Association for Machine Translation.

Declerck, Thierry, Thomas Troelsgård, and Sussi Olsen (2023c). "Towards an RDF Representation of the Infrastructure consisting in using Wordnets as a conceptual Interlingua between multilingual Sign Language Datasets." In Proceedings of the 12th Global Wordnet.

Declerck, Thierry, and Sussi Olsen (2023). "Linked Open Data compliant Representation of the Interlinking of Nordic Wordnets and Sign Language Data." In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pp. 62-69.

Declerck, Thierry (2022). Integration of Sign Language lexical Data in the OntoLex-Lemon Framework. IDS-Verlag.

Declerck, Thierry (2022). "Towards a new Ontology for Sign Languages." In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 3977-3983.

Declerck, Thierry (2022). Towards the Linking of a Sign Language Ontology with Lexical Data." In Proceedings of Globalex Workshop on Linked Lexicography within the 13th Language Resources and Evaluation Conference, pp. 6-9.

Garabík, Radovan, Ana Ostroški Anić, Sigita Rackevičienė, Giedrė Valūnaitė-Oleškevičienė, Linas Selmistraitis, and Andrius Utka (2023). "Validation of the Bigger Analogy Test Set Translation into Croatian, Lithuanian and Slovak." In Proceedings of the 4th Conference on Language, Data and Knowledge, pp. 402-409.

Garabík, Radovan, ed. (2022) LLOD Approaches for Language Data Research and Management: LLODREAM2022: International Scientific Interdisciplinary Conference, September 21-22, 2022: Abstract Book. Mykolas Romeris University.

https://llodapproaches2022.mruni.eu/wp-content/uploads/2022/10/LLOD_2022-Book-of-Abstracts.pdf

Gromann, Dagmar, Elena-Simona Apostol, Christian Chiarcos, Marco Cremaschi, Jorge Gracia, Katerina Gkirtzou, Chaya Liebeskind, Liudmila Mockiene, Michael Rosner, Ineke Schuurman, Gilles Sérasset, Purificação Silvano, Blerina Spahiu, Ciprian-Octavian Truică, Andrius Utka, and Giedre Valunaite Oleskeviciene (2024a). Multilinguality and LLOD: A survey across linguistic description levels. Semantic Web Journal, IOS Press. DOI: 10.3233/SW-243591

Gromann, Dagmar, Hugo Gonçalo Oliveira, Lucia Pitarch, Elena-Simona Apostol, Jordi Bernad, Eliot Bytyçi, Chiara Cantone, Francesca Frontini, Radovan Garabik, Jorge Gracia, Litizia Granata, Anas Fahad Khan, Timoteij Knez, Penny Labropoulou, Chaya Liebeskind, Maria di Buono, Ana Ostroški Anić, Sigita Rackevičienė, Ricardo Rodrigues, Gilles Sérasset, Linas Selmistraitis, Sidibé Mahammadou, Purificação Silvano, Blerina Spahiu, Enriketa Sogutlu, Ranka Stanković, Ciprian-Octavian Truică, Giedrė Valūnaitė Oleškevičienė, Slavko Zitnik, and Katerina Zdravkova (2024b). "MultiLexBATS: Multilingual Dataset of Lexical Semantic Relations." In Proceedings of LREC-COLING 2024. https://zenodo.org/records/10956565

Mishev, Kostadin, Ana Gjorgjevikj, Irena Vodenska, Lubomir T. Chitkushev, and Dimitar Trajanov (2020). "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers." IEEE access 8: 131662-131682.

Trajanov, Dimitar, Vangel Trajkovski, Makedonka Dimitrieva, Jovana Dobreva, Milos Jovanovik, Matej Klemen, Aleš Žagar, and Marko Robnik-Šikonja. (2023). "Review of Natural Language Processing in Pharmacology." *Pharmacological Reviews* 75, no. 4: 714-738.

Trajanov, Dimitar, Elena-Simona Apostol, Radovan Garabík, Katerina Gkirtzou, Dagmar Gromann, Chaya Liebeskind, Cosimo Palma, Michael Rosner, Alexia Sampri, Gilles Sérasset, Blerina Spahiu, Ciprian-Octavian Truică, and Giedre Valunaite Oleskeviciene (2024). "From Linguistic Linked Data to Big Data." In Proceedings of LREC-COLING 2024. https://zenodo.org/records/10956967

Rackevičienė, Sigita, Liudmila Mockienė, Andrius Utka, and Aivaras Rokas (2021). "Methodological Framework for the Development of an English-Lithuanian Cybersecurity Termbase." *Studies about Languages* 39, 85-92.

Rackevičienė, Sigita, Andrius Utka, Agnė Bielinskienė, and Aivaras Rokas (2022). "Distribution of Terms Across Genres in the Annotated Lithuanian Cybersecurity Corpus." *Respectus Philologicus* 41, no. 46, 26-42.

Rosner, Michael, Sina Ahmadi, Elena-Simona Apostol, Julia Bosque-Gil, Christian Chiarcos, Milan Dojchinovski, Katerina Gkirtzou et al. (2022). "Cross-Lingual Link Discovery for Under-Resourced Languages." In 13th International Conference on Language

Resources and Evaluation (LREC), JUN 20-25, 2022, Marseille, France, pp. 181-192. European Language Resources Association ASSOC-ELRA.

Rizinski, Maryan, Hristijan Peshov, Kostadin Mishev, Milos Jovanovik, and Dimitar Trajanov (2024). "Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (XLex)." IEEE Access .

Schuurman, Ineke, Thierry Declerck, Caro Brosens, Margot Janssens, Vincent Vandeghinste, and Bram Vanroy (2023). "Are there just WordNets or also SignNets?." In Proceedings of the 12th Global Wordnet Conference, pp. 172-178.

Other references in this report

Janev, Valentina, Damien Graux, Hajira Jabeen, and Emanuel Sallinger (2020a). Knowledge Graphs and Big Data Processing. Springer Nature.

Janev, Valentina Dea Pujić, Marko Jelić, and Maria-Esther Vidal (2020b). Chapter 9 Survey on Big Data Applications. In Valentina Janev, Damien Graux, Hajira Jabeen, and Emanuel Sallinger, editors, Knowledge Graphs and Big Data Processing, pages 149–164. Springer International Publishing, Cham.