Nexus Linguarum

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Deliverable 4.2

Intermediate Activity Report. Working Group 4 “Use Cases and Applications”

Main authors: Sara Carvalho, Ilan Kernerman

Contributors: Kristina Despot, Slavko Žitnik, Barbara Lewandowska-Tomaszczyk, Gordana Hržica, Mietta Lennes, Jouni Tuominen, Florentina Armaselu, Mariana Damova, Valentina Janev, Daniela Gifu, Sigita Rackevičienė, Dimitar Trajanov, Ana Ostroški Anić, Marko Robnik-Šikonja, Petya Osenova, Barbara McGillivray

30 October 2021

Project Acronym	NexusLinguarum
Project Title	European network for Web-centred linguistic data science
COST Action	18209
Starting Date	26 October 2019
Duration	48 months
Project Website	https://nexuslinguarum.eu/
Chair	Jorge Gracia
Main authors	Sara Carvalho and Ilan Kernerman
Contributors	Florentina Armaselu, Mariana Damova, Kristina Despot, Daniela Gifu, Gordana Hržica, Valentina Janev, Mietta Lennes, Barbara Lewandowska-Tomaszczyk, Barbara McGillivray, Petya Osenova, Ana Ostroški Anić, Sigita Rackevičienė, Marko Robnik-Šikonja, Dimitar Trajanov, Jouni Tuominen, Slavko Žitnik
Reviewers	NexusLinguarum core group team
Version | Status	final
Date	30 October 2021

Acronyms List

CA	COST Action
ICT	Information and Communication Technologies
GP	Grant Period
LD	Linked Data
LLD	Linguistic Linked Data
LLOD	Linguistic Linked Open Data
LOD	Linked Open Data
LR	language resource
ML	Machine Learning
NL	NexusLinguarum
NLP	Natural Language Processing
SA	Sentiment Analysis
SALLD	Sentiment Analysis and Linguistic Linked Data
SOTA	state-of-the-art
STSM	Short Term Scientific Mission
UC	Use Case
WG	Working Group
WSD	word sense disambiguation

Table of Contents

Executive Summary

This report focuses on the evolution of the tasks and use cases explored in Working Group 4 (WG4) of the NexusLinguarum COST Action (CA 18209), especially over the last 6 months (May-October 2021). It builds upon the first deliverable (D4.1 – Use Case Description and Requirements Elicitation), submitted in April 2021, and a journal article published in July 2021 (*Lexicala Review*, 29: 26-72), both of which comprising a comprehensive description of the use cases, the requirements necessary for their implementation, and their achievements. In addition to the progress of each of the 4 tasks and, in particular, of the 8 use cases, this intermediate report elicits the increasingly solid inter-WG cooperation and WG4's latest activities, as well as next steps, mainly regarding the upcoming Grant Period (GP3, November 2021 – October 2022).

1. Introduction

Working Group 4 of the NexusLinguarum CA explores use cases and applications in which the Action's relevant methodologies, technologies, and standards can be tested and validated. Since April 2021, the number of members has stabilised (approximately 110), and their multidisciplinary backgrounds remain one of WG4's biggest assets. There has been a very active participation by some of these members in the nine WG bimonthly meetings held thus far.

The structure underpinning the WG incorporates this interdisciplinary approach and benefits from having two leaders for each Task, with backgrounds in Linguistics and Computer Science, respectively. Given the wide range of selected domains (Linguistics, Humanities and Social Sciences, Technology, and Life Sciences), a second level has been created, where the actual use cases and applications are developed. The current structure is depicted in Figure 1, with each of the four Tasks incorporating two Use Cases (UCs). Although a call for new UCs was launched in M17, no proposals have been received so far, so the structure below, although not closed, appears to be stable when the CA is now reaching half term.

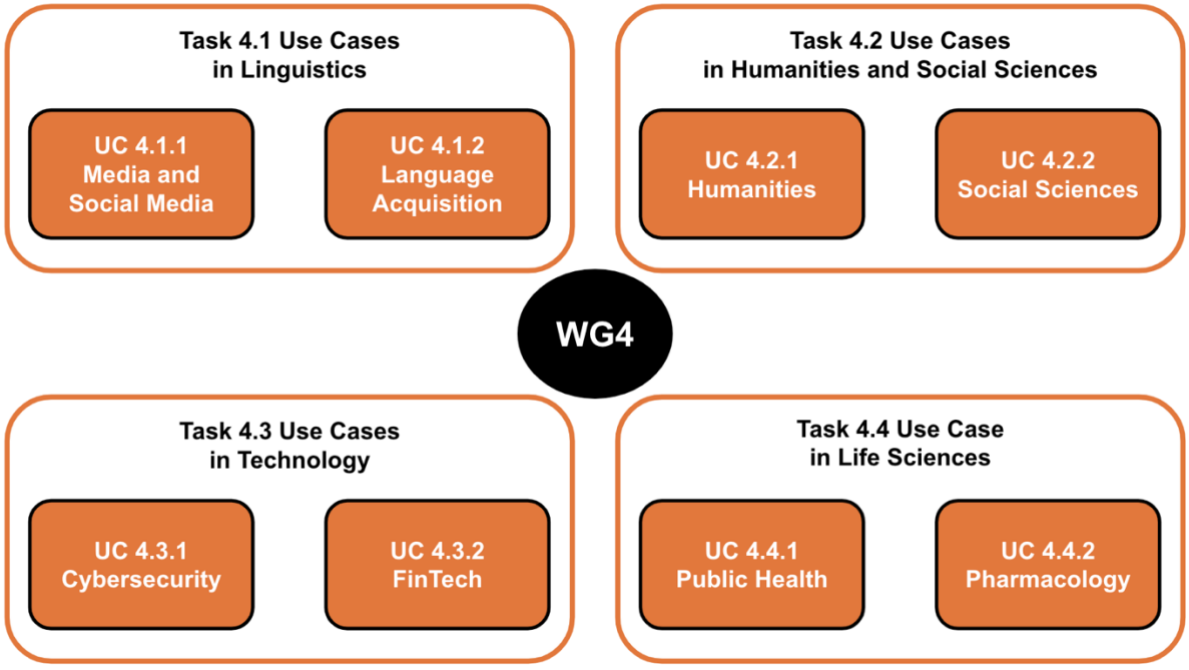

Figure 1. WG4's structure of Tasks and UCs (as of October 2021)

The WG's core team has also remained solid and has been quite actively engaged throughout these first 24 months, with just a few changes to report, as outlined in Table 1:

ROLE	PERSON	COUNTRY
WG4 leader	Sara Carvalho	Portugal
WG4 co-leader	Ilan Kernerman	Israel
T4.1 leader – linguistics	Kristina Despot	Croatia
T4.1 leader – computational	Slavko Žitnik	Slovenia
UC4.1.1 coordinator	Barbara Lewandowska-Tomaszczyk	Poland
UC4.1.2 coordinator	Gordana Hrzica	Croatia
T4.2 leader – linguistics	*Ana Luís (until May 2021) Mietta Lennes (from June 2021)	Portugal Finland
T4.2 leader – computational	Jouni Tuominen	Finland
UC4.2.1 coordinator	Florentina Armaselu	Luxembourg
UC4.2.2 coordinator	Mariana Damova	Bulgaria
T4.3 leader – linguistics	Daniela Gifu	Romania
T4.3 leader – computational	*Valentina Janev (until October 2021) Dimitar Trajanov (from November 2021)	Serbia North Macedonia
UC 4.3.1 coordinator	Sigita Rackeviciene	Lithuania
UC4.3.2 coordinator	Dimitar Trajanov	North Macedonia
T4.4 leader – linguistics	*Petya Osenova (until June 2021) Ana Ostroški Anić (from June 2021)	Bulgaria Croatia
T4.4 leader – computational	Marko Robnik-Šikonja	Slovenia
UC4.4.1 coordinators	Petya Osenova Marko Robnik-Šikonja	Bulgaria Slovenia
UC4.4.2 coordinator	Dimitar Trajanov	North Macedonia

Table 1. WG4's core team

At the onset of the CA, and especially throughout the first Grant Period (GP1, October 2019 – April 2020), the main goals of the WG consisted of: i) selecting the relevant tasks and use cases; ii) devising the WG structure, core team, and workflow; and iii) preparing an initial description of each task and UC. In the second GP (GP2, May 2020 – October 2021), the main focus was on the UC definitions and requirements' elicitation, both of which were thoroughly reported in the first deliverable, submitted in April 2021, as well as in a journal article published in *Lexicala Review* in July 2021 (Carvalho and Kernerman 2021). The actual work on the various UCs, whose main updates are further explored in Section 2, was also intensified during GP2. Moreover, close collaboration with the remaining NexusLinguarum WGs, a critical part of WG4's mission, continued to be actively fostered and enhanced over the last year, as shown in Section 3.

2. Tasks and Use Cases

2.1. Task 4.1. Use Cases in Linguistics

Task Leaders Kristina Despot (linguistics), Slavko Žitnik (computational)

Use Cases

UC 4.1.1 Media and Social Media

UC 4.1.2 Language Acquisition

Status update and next steps

The task consists of two use cases focusing on Media and Social Media (UC4.1.1) and Language Acquisition (UC4.1.2). The first deals mostly with offensive language analysis from the linguistic perspective, applying text analytics tools to support linguistic phenomena. The second is focused on the development of tools for language acquisition and analysis.

Both use cases operate in parallel and small focus groups are formed to work on specific tasks. We organize lectures regarding specific topics where new ideas, collaborations and possible future directions are discussed. The task leaders also organize overall meetings, particularly after specific tasks are concluded, to reorganize focus groups and continue with the next steps.

2.1.1. UC 4.1.1. Use Case in Media and Social Media

Coordinator Barbara Lewandowska-Tomaszczyk

Status update

The Use Case in Incivility in Media and Social Media (UC4.1.1) is involved in researching offensive language, as well as exploring its recognition and identification methods in everyday texts, mostly in media and social media, which might contain abusive content. Offensive language research may lead to establishing algorithms that spot offensive content and enable automatic protection of users from undesirable messages. In the past several months, we have been researching the existing corpora of English. We have used exploratory techniques leveraging both pure linguistic knowledge (i.e. theory, SketchEngine[1] corpora and tools) and computational linguistic processing of available data (i.e. (non-)contextual embeddings and BERT). As an intermediate result, we proposed a taxonomy of offensive language and submitted two papers for publication:

[1] https://www.sketchengine.eu/

1. Lewandowska-Tomaszczyk, B., S. Žitnik, A. Bączkowska, Ch. Liebeskind, J. Mitrovic, G, Valuntaite. (submitted to SALLD Proc.) LOD-connected offensive language Ontology and Tagset Enrichment.

2. Žitnik,S., B. Lewandowska-Tomaszczyk, A. Bączkowska, Ch, Liebeskind, G. Valuntaite, J. Mitrovic, *Detecting Offensive Language: A New Approach to Offensive Language Data Preparation* (submitted to Natural Language Engineering).

On 28 September 2021, we organized a one-day Workshop on *Abusive Data Annotation* at the Faculty of Computational Linguistics and Engineering of St Cyril and Methodius University in Skopje. The hybrid workshop, which attracted over 30 participants, was organized in liaison with the NexusLinguarum plenary meeting.

Next steps

There are two avenues of our next research. The first task concerns extending our investigation to Offensive Implicit Language categorization and verification relying on both linguistic knowledge and adequate computational tools. The second task refers to the verification of the proposed taxonomy by annotating offensive language datasets in different languages.

To perform these objectives, cooperation plans are being put forward to boost our inter-WG cooperation in the NexusLinguarum CA, in particular with regard to research presented in WG3. Furthermore, in the Spring of 2022 we plan to organize a workshop on *Explicitness and Implicitness in Offensive Language* at the Jerusalem College of Technology, also from the perspective of LLOD, envisaged as a task for the last year of the Action lifetime.

2.1.2. UC 4.1.2. Use Case on Language Acquisition

Coordinator Gordana Hržica

Status update

This use case has two foci: the first focal point concerns the development of tools based on language technologies to be used in language assessment, a process conducted by speech and language therapists as a part of a diagnostic procedure. In language assessment, analysis of spontaneous or elicited discourse production is often neglected because it is timely and requires a lot of background linguistic knowledge. By developing tools that foster the analysis, we contribute to the recognition of children and adults with language disorders.

Other practitioners, such as teachers of first and second language or psychologists, can also use the app to follow the progress of children, for instance. Currently, the app is being developed for Croatian language and presented to scientists and practitioners in Croatia and abroad (Figure 2). We held three presentations in two conferences and in one workshop.

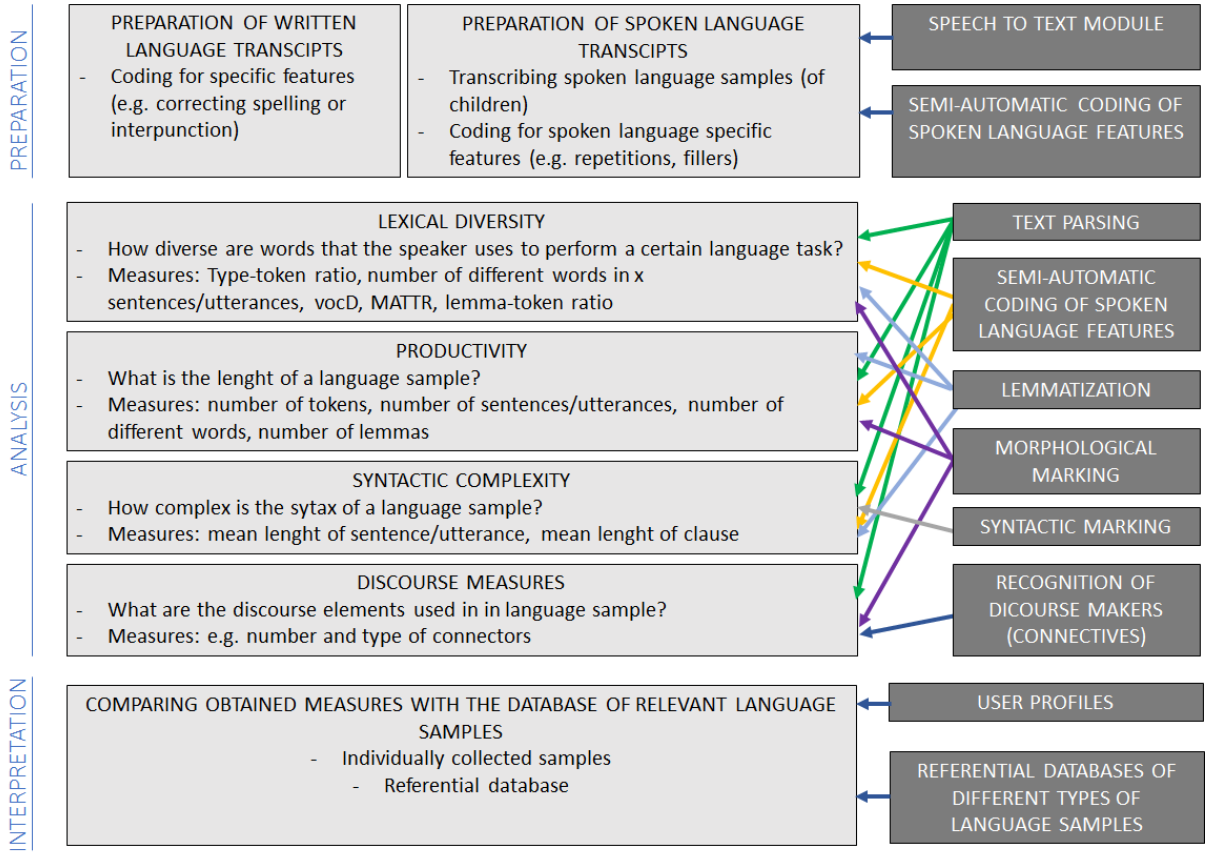

Figure 2: *MultiDis Application: Overview*

The second topic concerns first language acquisition research using digital language resources and language technologies. We are currently exploring three research questions: (1) What is the relation between lexical diversity and morphological complexity having in mind cross linguistic differences? (2) Can we establish a correlation between lexical complexity and the chronological age of a child? (3) What is the role of figurative speech in early child language?

These questions have not been broadly explored in crosslinguistic research. The analyses are based on the existing corpora (Child Language Exchange System - CHILDES) database, which is the largest database of child language – mostly spoken language – and which is part of the TalkBank (MacWhinney, 2002). Three papers are in progress, the one dedicated to research question 1 being the most advanced.

For this paper, we used diverse measures of lexical diversity and morphological complexity (mean average type-token radio, vocabulary D, word entropy, relative entropy of the word structure, lexical complexity). We analysed children's narratives obtained by the same wordless storybook. Comparable groups of children speaking one of the eight languages were included in the analysis. This has allowed us to compare the relationship between measures of lexical diversity and morphological complexity in different languages.

On the other hand, the paper devoted to research question 2 explores alternative methods for establishing lexical values, namely subjective measures obtained by native speakers of certain languages, and their correlation to age of acquisition. Lastly, the paper dealing with research question 3 focuses on qualitative analysis of figurative speech, which may later be used to foster computational analysis of figurative speech in early language development.

Next steps

In the following months, the group will work on the following tasks regarding the MultiDis application (cf. Figure 3): (1) Implementation of the automatic calculation of language development measures; (2) Making the referential database of transcripts ready for individual and group comparisons; (3) Implementing and testing Speech-to-text module. At the same time, we will work on the MultiDis user interface, aiming to simplify the process of analysis. We will continue to promote the app among scientists and professionals working in the field.

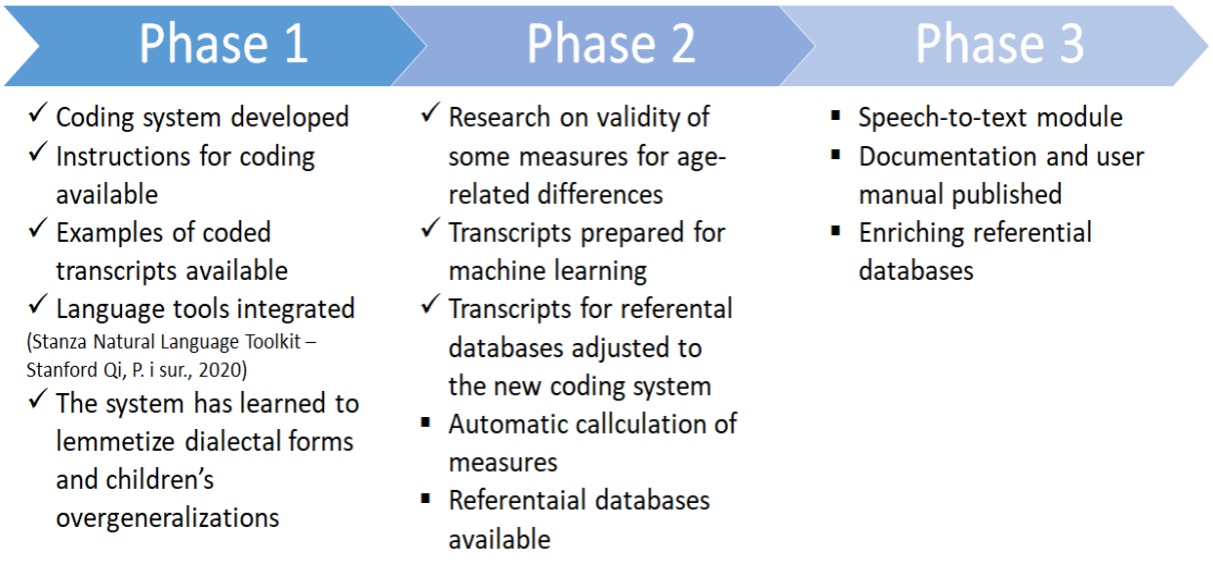

Figure 3: *Application MultiDis: Current status and future plans*

Regarding our publications, by the end of 2021 we would like to submit the paper about lexical diversity and morphological complexity, and concentrate heavily on the other two papers, to be ready in the first months of 2022. At the same time, we will submit papers for conferences. We will start preparing the paper about the MultiDis application to be submitted to a methodologically oriented journal.

In the near future, we would like to explore possible collaborations with other Working Groups, following suggestions and guidelines provided by other NL members during the Skopje meeting. We also hope to expand this UC core team, if possible, with researchers interested in the adaptation of the MultiDis app to their own respective languages.

2.2. Task 4.2. Use Cases in Humanities and Social Sciences

Task Leaders Mietta Lennes (linguistics), Jouni Tuominen (computational)

Use Cases

UC 4.2.1 Humanities

UC 4.2.2 Social Sciences

Status update and next steps

Task 4.2 comprises two use cases, focusing on Humanities (UC4.2.1) and Social Sciences (UC4.2.2). UC4.2.1 explores how linguistic data science can be used for studying the evolution of (parallel) concepts in various languages and fields of the Humanities. UC4.2.2 investigates the use and development of language processing tools that facilitate the usage of survey data archives.

Having surveyed the state-of-the-art and identified the corpora to be used, these use cases have advanced to the phases of conceptualization of the methodology for the identification and modelling of semantic change (UC4.2.1) and method development for detection, extraction and semantic classification of discourse markers (UC4.2.2). The use case coordinators have organized activities, such as joint publications and STSMs, as well as searched and applied for funding opportunities regarding future research projects. Some collaboration possibilities with other WGs have already been identified, and we will continue to explore others.

2.2.1. UC 4.2.1. Use Case in Humanities

Coordinator Florentina Armaselu

Status update

The use case combines NLP and Semantic Web technologies to trace the history of concepts in the domain of socio-cultural transformation. The focus of this study is on the detection and representation of semantic change using multilingual diachronic corpora and exploring techniques such as diachronic word embedding and ontology learning from text (e.g. the "ontology learning layer cake" model and methodology for the acquisition of *terms, synonyms, concepts, concept hierarchies, relations* and *rules*). By semantic change, we understand a change in meaning, either of a lexical unit (word or expression) or of a concept (a complex knowledge structure that can encompass one or more lexical units, as well as relations among them and with other concepts).

UC 4.2.1 is currently in the phase of conceptualisation intended to define the methodology to be applied in the project. This represents an intermediary stage, in-between the examination of the state-of-the-art in various domains relevant to the fields of research targeted in the study, and the implementation of the designed methods. So far, we have identified the datasets to be processed, and defined the general workflow (and specific areas of interest within the constituent blocks), as illustrated in the figure below (Figure 4).

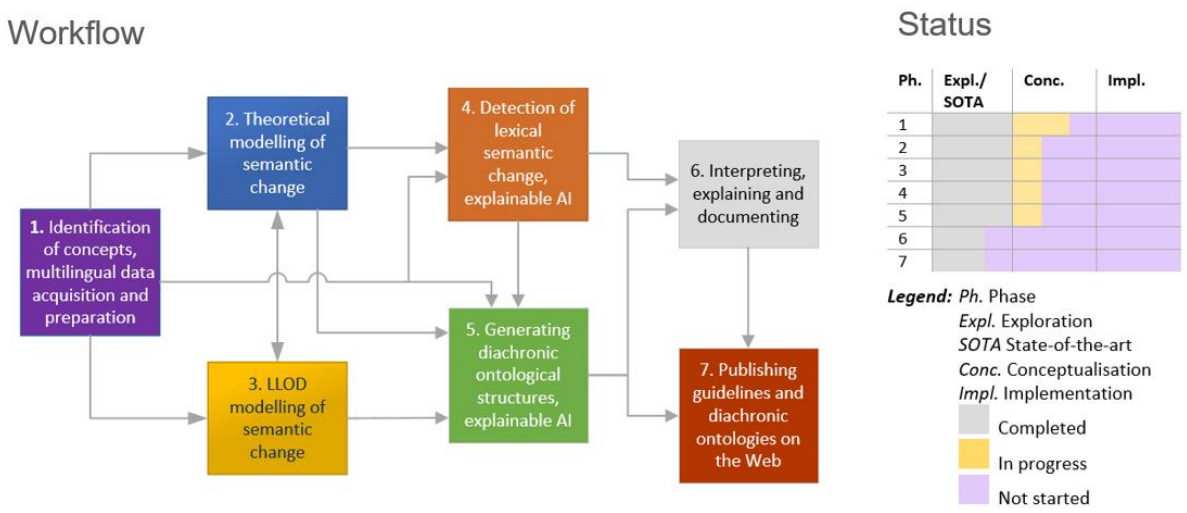

Figure 4: *UC4.2.1 workflow*

The multilingual diachronic corpora to be used in the project include a core dataset that may be extended to other languages, through collaboration with other WG4 use cases:

- Core dataset (literary, religious, technical, philosophical, historiographical, everyday life matters, newspapers, etc.): LatinISE (**Latin**, 2nd c. BC - 20th c.CE); Diorisis (**Ancient Greek,** 8th c. BCE - 5th c. CE); Responsa (**Hebrew,** 11th - 21st c.); Sliekkas (**Lithuanian,** 16th - 18th c.); BnL Open Data collection (**French, German, Luxembourgish**, 1690-1918, 1841-1878).
- Extended dataset: **Bulgarian** (19th - 20th c.); **English** (15th - 21st c., 2011-2018); **Polish** (16th -18th c., 1945–1962); **Romanian** (1817-2013); **Slovene** (1584-1919); **multilingual** (2001-2021, 2001 – 2012).

UC 4.2.1 outcomes:

- Publication: Armaselu, Florentina. Apostol, Elena-Simona. Khan, Anas Fahad. Liebeskind, Chaya. McGillivray, Barbara. Truică, Ciprian-Octavian. Valūnaitė Oleškevičienė, Giedrė. "HISTORIAE, History of Socio-Cultural Transformation as Linguistic Data Science. A Humanities Use Case", *Open Access Series in Informatics (OASIcs), 3rd Conference on Language, Data and Knowledge (LDK 2021)*, Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, Barbara Heinisch (eds.), Volume 93, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2021.
- Submitted paper (under review): Armaselu, Florentina. Apostol, Elena-Simona. Khan, Anas Fahad. Liebeskind, Chaya. McGillivray, Barbara. Truică, Ciprian-Octavian. Utka, Andrius. Valūnaitė Oleškevičienė, Giedrė. Van Erp, Marieke. "LL(O)D and NLP Perspectives on Semantic Change for Humanities Research", *Semantic Web Journal*, CFP: Special Issue on Latest Advancements in Linguistic Linked Data, 2021.
- Submitted funding application (not retained): HIstory of Socio-culTural transfORmation as linguistIc dAta sciEnce (HISTORIAE). CHANSE, Call for Proposals "Transformations: Social and Cultural Dynamics in the Digital Age", Outline Proposal, 2021.

Next steps

In the coming months, the group will work on the following tasks: (1) preparation of the datasets for diachronic analysis and knowledge extraction (conversions, metadata structure and acquisition, organizing the collections by time slice, year, decade, centuries, etc.); and design of the (2) theoretical and (3) LLOD model for capturing semantic change. The second and third tasks will address questions related to the modelling of semantic change pertaining to phenomena defined in lexical semantics, conceptual history, and knowledge organization, such as semasiological and onomasiological innovation, and concept drift, as well as the representation of space-time from a diachronic perspective through LLOD and other Semantic Web formalisms.

Further development will include (4) applying deep learning methods for lexical semantic change detection and (5) diachronic ontology learning, in combination with explainable AI techniques, and the publication of a (6) set of guidelines resulting from the use case and (7) sample of diachronic ontologies on the LLOD cloud. Points of potential collaboration with colleagues from WG1, 2 and 3 have already been identified (see the inter-WG matrix) and will be further discussed and analysed in the near future through use case and inter-WG meetings, and other types of joint activities (publications, workshops, etc.).

2.2.2. UC 4.2.2. Use Case in Social Sciences

Coordinator Mariana Damova

Status update

This UC focuses on building a toolset of language processing tools that enable the linguistic analysis of survey data. It aims at exploring language phenomena that encode speakers' attitudes, as well as researching and developing methods that help discover, classify and represent them as Linguistic Linked Open Data (LLOD). Studying survey data, social media data and other texts charged with attitudes, such as public speeches, will result in generalizations about social attitudes clusters that will also be classified and represented as models.

In particular, we work on TED talks and discourse markers, and have constituted a discourse marker vocabulary of 383 multiword expressions that may or may not occur in text as discourse markers. We have constituted a parallel corpus in 6 languages – English, Latvian, Bulgarian, Portuguese, German and Macedonian - including 44.5K sentence contexts with English as a pivot language. In addition, we have manually annotated 2,340 sentence chunks in English, Bulgarian and Lithuanian for the availability of a discourse marker.

Afterwards, we have run transformer ML models trained with these data to predict the availability of discourse markers in unseen English and Bulgarian contexts, with the best prediction score reaching 95% precision. We have also generated predictions for the availability of discourse markers in all the languages covered in the parallel corpus, based on the model for English. With respect to semantics and LLOD, we studied theoretical frameworks about the semantics of discourse markers and the discourse relations they trigger, e.g., RST (rhetorical structure theory), PDTB (Penn Discourse Treebank), Crible (annotation protocol for spoken corpora), SDRT (segmented discourse representation theory), as well as LLOD vocabularies such as OLiA, LEMON, rdf4discourse, lexicon of discourse markers for Portuguese. For this latter part, we liaised with WG1.

We have selected the ISO 24617-8:2016 standard for representing discourse relations in order to semantically annotate the semantic contribution of discourse markers in text. Furthermore, we designed a template for manual annotation according to that ISO standard. In addition, we have carried out two STSMs within the Call "Attitudes Detection and Representation in Survey Data" (*Researching discourse markers expressing opinion with machine learning techniques in a multilingual corpus* and *Machine learning for detecting and interpreting language phenomena in survey data*) and presented a paper at the *Deep Learning and Neural Approaches for Linguistic Data* Workshop in September 2021, in Skopje.

Next steps

We plan to: 1) verify and validate the results generated by ML predictions for the availability of discourse markers in text; 2) to extend the ISO annotation schema with pragmatic factors distinguishing opinion and attitude pointers, including offensive language, among discourse markers, and 3) to organize manual annotation for the 6 languages of the parallel corpus. Consequently, we will apply Transformer ML models to predict semantics of discourse relations and constitute a corpus of survey data, in order to apply the created predictive models to survey texts, for example on COVID-19 survey data of the European Parliament. We will extend the initiative to other languages, like Polish and Spanish, to add multimodality considerations by investigating the intonation in expressing discourse markers in speech, which will liaise us with WG3. Moreover, we also plan to study disambiguation techniques, thereby liaising with WG2. In addition, it is possible to consider participation in Discourse Relation Parsing and Treebanking (DISRPT) and the submission of a full paper to the *Computational Linguistics* Journal. Our main focus in GP3 will be looking for funding opportunities to submit a full research proposal that will allow us to conduct our initiative in more depth.

2.3. Task 4.3. Use Cases in Technology

Task Leaders Daniela Gifu (linguistics), Valentina Janev (computational)

Use Cases

UC 4.3.1 Cybersecurity

UC 4.3.2 Fintech

Status update and next steps

Task 4.3 builds upon the recent advancements in the areas of multilingual technologies, machine translation, automatic term extraction methods, text analytics and sentiment analysis models, with the aim to reuse existing open-source components and test them in different ICT and business scenarios.

During the first year of the CA, two specific Use Cases have been selected:

- Cybersecurity - the goal of the Cybersecurity UC is the compilation of a bilingual/multilingual termbase of cybersecurity terms and their metadata in at least 3 languages.
- FinTech - the goal of the FinTech UC is to develop domain-specific sentiment analysis models that can provide an efficient method for extracting actionable signals from the news.

In the second year of the Action, both use cases made significant progress, resulting in publications further specified below.

Moreover, T4.3 leadership was involved in the local organization of the 15th EUROLAN Training School, dedicated to the *Introduction to Linked Data for Linguistics*. This online event, which was supported and coordinated by NexusLinguarum, was held on February 8-12, 2021, and was attended by 82 participants. Further details are available in the dedicated report.

2.3.1. UC 4.3.1. Use Case in Cybersecurity

Coordinator: Sigita Rackevičienė

Status update

In the period September 2020 – October 2021, the following tasks have either been completed or are at the final completion stage:

T1. Research on the existing Cybersecurity (CS) terminology resources (analysis of existing termbases and glossaries, systematisation of the collected information – completed).

T2. Compilation of English-Lithuanian CS corpora:

T2.1. The corpus is composed of EU documents on cybersecurity issues, including 1,408,736 words (final completion stage).

T2.2. Comparable CS corpus composed of international and national documents (legal acts, reports and recommendations of CS agencies, academic literature, mass and specialised media articles), including 4,002,918 words (final completion stage).

T2.3. The gold standard CS corpora with manually labelled CS terms for training neural network systems (final completion stage):

- parallel gold standard corpus – 101,150 words (currently 3,361 tagged EN terms and 3,348 tagged LT terms);
- comparable gold standard corpus– 164,238 words (currently 3,223 tagged EN terms and 8,158 tagged LT terms).

The current structures of the corpora are presented in Figure 5 below (the gold standard corpora have been composed according to the structure of the large-scale corpora):

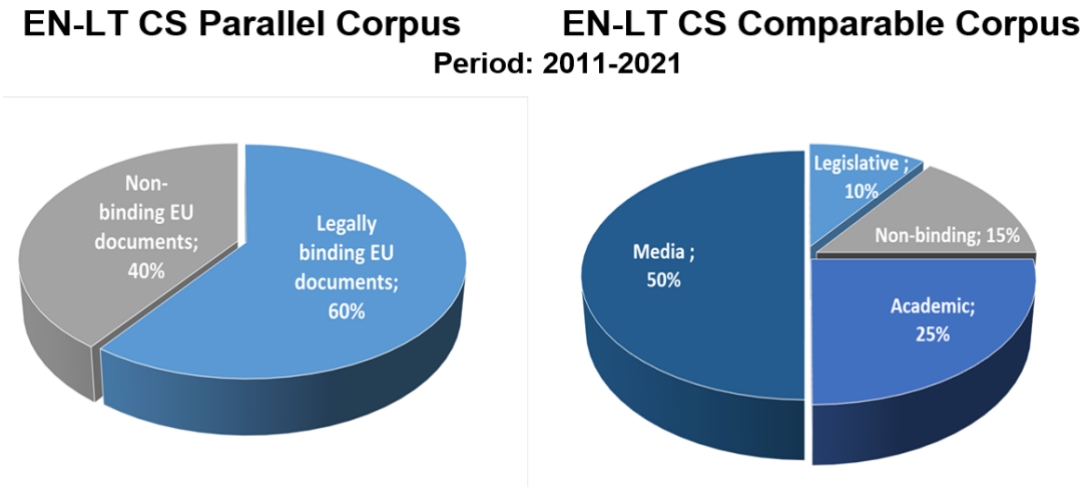

Figure 5: *UC 4.3.1 corpora*

T3. Automatic extraction of terminological data and metadata

The pilot study on extraction of Lithuanian CS terminology using a small-scale annotated dataset (completed). The results were published in: Rokas, Aivaras; Rackevičienė, Sigita; Utka, Andrius. "Automatic Extraction of Lithuanian Cybersecurity Terms Using Deep Learning Approaches". In *Human Language Technologies – The Baltic Perspective. Proceedings of the Ninth International Conference Baltic HLT 2020*. IOS Press, 2020.

Next steps

- Completion of tasks T2.1, T2.2 and T2.3.
- **Task T3. Automatic extraction of terminological data and metadata**: T3.1. Development of automatic bilingual terminology extraction methods by training and iterative testing of the neural networks using the gold standard datasets; T3.2. Selection of the most effective methods and automatic extraction of term candidates, their automatic alignment; T3.3. Selection of the dominant CS terms based on frequency/dispersion analysis and expert approval; T3.4. Development of automatic methods for extraction of knowledge-rich contexts and their extraction for the selected dominant CS terms.
- **T4. Compilation of the termbase**: T4.1. Formulating final definitions of the terms using the extracted knowledge-rich contexts; T4.2. Collecting other metadata about the selected terms: usage examples; conceptual relations with other terms, statistical data on term usage; T4.3. Uploading the collected data and metadata to a termbase.
- **T5. Interlinking the termbase with other resources and its application:** interlinking the bilingual termbase data with other resources.
- **T6. Analysis of the conceptual, linguistic, and pragmatic dimensions of the collected terminology**: conceptual categorisation of CS terminology, analysis of its origin, structural patterns and usage in different genres (density and diversity of terminology, usage of initialisms, etc.).

2.3.2. UC 4.3.2. Use Case in Fintech

Coordinator Dimitar Trajanov

Status update

A comprehensive chronological study of NLP-based methods for sentiment analysis in finance was made and more than 100 models for sentiment analysis were analysed. The research started with a lexicon-based approach, then moved to word and sentence encoders, and finally to contemporary NLP transformers. In comparison to the other approaches examined, the NLP transformers perform exceptionally well. The text representation models, which input the semantic meaning of words and phrases into the models, are responsible for most of the improvement in the sentiment analysis accuracy. The top models produce outcomes that are equivalent to expert judgment. The assessments were carried out on a relatively small dataset of around 2,000 phrases.

Despite the small size of the dataset, we achieved excellent results, indicating that this approach is acceptable for domains where big datasets are not available. The analysis has shown that the distilled versions (Distilled-BERT and Distilled-RoBERTa) achieve text classification performances comparable to their large, uncompressed teacher models. As a result, they can be utilized efficiently in text classification production environments, where the need for lightweight, responsive, energy-efficient, and cost-saving models is essential. The findings of this work have applications in fields such as finance, where decision-making is dependent on sentiment extraction from large textual databases.

The results suggest that some models may be successfully utilized to project stock market trends and corporate profitability, make securities trading and portfolio management decisions, as well as manage brand reputation. Although this method was designed for sentiment analysis in finance, it may be used in other fields such as healthcare, legal, and business analytics.

Next steps

For the next period, we have two goals: to apply sentiment analysis results in a variety of areas in finance, and to utilize Explainable AI (XAI) to reverse-engineer the sentiment analysis models. In the application area, we plan to start building models for Cryptocurrency price prediction using the sentiment of news and social networks. The advancement in Explainable AI models inspires their application to sentiment analysis in finance. We will start with the creation of XAI on top of the transformer-based models that have the best performance. This will allow us to understand the connections between the text characteristics and the resulting sentiment. We plan to create an algorithm for the automatic generation of a FinTech sentiment dictionary based on this analysis.

2.4. Task 4.4. Use Cases in Life Sciences

Task Leaders Ana Ostroški Anić (linguistics) and Marko Robnik-Šikonja (computational)

Use Cases

UC 4.4.1 Public Health

UC 4.4.2 Pharmacology

Status update and next steps

The use of text analytics in life sciences is extensive. Still, it primarily focuses on English language and information retrieval, intelligent search and creation of linked data (like knowledge graphs). In this task, we address less researched areas of linguistic data science, in particular public health and pharmacology, and cover several less-resourced languages.

The public health use case (UC 4.4.1) has recently become the primary focus of global research due to the COVID-19 epidemics, while the use case in pharmacology (UC 4.4.2) addresses a well-established area of broad public interest. In both use cases, our research has moved beyond initial investigation (description of research areas, main resources, and methodology), and we now work on specific research tasks specified below.

In UC 4.4.1, we did an initial cross-lingual analysis of COVID-19 related topics and metaphors, focusing on parliamentary data. In UC 4.4.2, we created a predictive model to discover and label the drug-disease relations from scientific texts.

In the next phase, we will extend the set of covered languages, analyzed datasets and analytical approaches. In UC 4.4.1, we will compare the results of metaphoric analysis on several languages and improve and generalize the initial ontological model based on feedback from other languages. In UC 4.1.2, we will attempt to predict diseases from electronic health records (EHR).

2.4.1. UC 4.4.1. Use Case in Public Health

Coordinators Petya Osenova and Marko Robnik-Šikonja

Status update

Three lines of research were executed:

1. Investigating, selecting and summarizing the existing literature, resources and methods related to Public Health.

2. Cross-lingual neural-based methods were applied to ParlaMint corpora of Bulgarian and Slovenian for detecting similarities/differences in pre-Covid and during-Covid parliaments.

The initial results showed that while in the Slovenian Parliament the during-Covid topics were related to health issues, in the Bulgarian Parliament no substantial changes were observed in pre-Covid and during-Covid topics.

3. Together with the colleagues in UC 4.1.1 (Media and Social Media), an investigation started of COVID-19 related metaphors in parliamentary discourse. Again, the ParlaMint corpora were used.

The annotation model was taken from the following paper where the domain was news media: Kristina Štrkalj Despot and Ana Ostroški Anić 2021: A War on War Metaphor: Metaphorical Framings in Croatian Discourse on Covid-19. Rasprave Instituta za Hrvatski Jezik i Jezikoslovlje, 47(1), 2021.

The first metaphor classification for parliamentary discourse was made for Lithuanian by Liudmila Mockienė within her STSM stay at IICT-BAS (Sofia, Bulgaria) in the period of 16-30 August 2021. She extracted the concordance contexts of the respective keywords and classified the examples into metaphorical/non-metaphorical usages. Then, the metaphorical usages were further classified according to the schema in the above-mentioned paper. The same survey was performed later for Bulgarian. A preliminary ontological formalization of COVID-related metaphor frames in parliamentary data was initiated with a mapping to the Lemon lexicography module for ontologies.

Next steps

1. Improving the cross-lingual methods of topic modelling on Slovenian and Bulgarian parliamentary corpora. Performing more experiments on the data with the respective quantitative and qualitative analysis. Adding more languages in the cross-lingual experiments as well as more data (covering not only the period of November 2019 to July 2020 but also from July 2020 up to December 2021) when it is ready from the upcoming next release of ParlaMint corpora.

2. Extending the metaphor frame classification to the Croatian, Slovene and Polish parliamentary corpora.

3. Validation of the initial ontological model over the metaphor frames. Discussions on which content parts related to metaphors will be modelled and to what extent. Expanding this model with the upcoming frames for the new languages. Here we envisage active collaboration with WG1.

2.4.2. UC 4.4.2. Use Case in Pharmacology

Coordinator Dimitar Trajanov

Status update

Our main objective was to discover ways of how natural language processing (NLP) is used in the field of Pharmacy. We analyzed more than 60 pharmacy-related papers that mention or utilize NLP models. The goal was to identify:

1. The pharmacy-related areas where NLP can be applied
2. Used NLP tasks and algorithm
3. Used dataset and data sources
4. NLP libraries related to the application of NLP in Pharmacy
5. Analysis of the application of linked data and semantic web related technologies in Pharmacy

1. The applications can be categorized into four main domains:

- Drug-drug interaction
- ADE – adverse drug events
- Pharmacovigilance
- Standardization of Drug information

2. Within the four observed domains, the most common NLP tasks used were classification and named entity recognition. In addition, multiple studies had mentioned the extraction of information from different sources and classifying the extracted information.

3. Some of the analyzed articles used a closed dataset, but there were a dozen publicly available datasets that can be reused in future research. We created a short description of the nature of the data, size, and possible usage for all these datasets.

4. As a fourth sub-task, we created a list of NLP libraries used in the pharmacy-related papers. Based on our experience, we also added an additional list of state-of-the-art NLP libraries that were not mentioned in the papers but can be applied in Pharmacy.

5. The last sub-task is surveying the application of semantic web and linked data-related technologies usage in Pharmacy.

Next steps

We plan to finish the survey and work on the application of NLP to Electronic Health Records-related data and texts. In addition, we intend to develop an NLP-based model for drug-disease relation detection and labelling, which will use a series of standard state-of-the-art models to evaluate a corpus of abstracts from scientific biomedical research publications.

3. Related activities

3.1. Interaction with the other Working Groups

Inter-WG collaboration, one of the axes underpinning this CA, has become increasingly visible, not only through regular participation of several CA members in dedicated WG and UC meetings, but also through joint publications and event organization. To help support these interactions and make the potential connections between WGs and tasks more explicit, WG4 proposed the creation of a matrix at the NexusLinguarum 2nd plenary meeting in Lisbon (October 2020) (cf. Carvalho and Kernerman 2021), which was then expanded at the 3rd plenary meeting in Skopje (September 2021). This extended version of the matrix reflects the discussions which have been taking place between the WGs during GP2, but which will necessarily be further developed throughout GP3. The figures below represent, therefore, the current stage of this inter-WG collaboration and will be updated accordingly in upcoming deliverables.

Figure 6 depicts the points of contact between WG1 and WG4, the prominence of Modelling (T1.1) and Interlinking (T1.3) being explained by the need to address the initial and current requirements/challenges identified by the various Use Cases within WG4. However, it should be noted that Quality (T1.4), metadata, and versioning (T1.2) have also been identified as relevant cross-UC topics.

	Task 4.1 Linguistics		Task 4.2 Humanities & Social Sciences		Task 4.3 Technology		Task 4.4 Life Sciences	
	UC4.1.1 Media and Social Media	UC 4.1.2 Language Acquisition	UC 4.2.1 Humanities	UC 4.2.2 Social Sciences	UC 4.3.1 Cybersecurity	UC 4.3.2 FinTech	UC 4.4.1 Public Health	UC 4.4.2 Pharmacy
T1.1 Modelling	LOD cross-linguistic modeling of hatespeech/offensive language taxonomies,levels and types.	? LLOD modelling of language acquisition data/tools, e.g., CHILDES corpus or DTA Tool	LLOD modelling of semantic change in diachronic corpora	LLOD modelling of discourse annotations and discourse marker inventories; semantics and pragmatics of speaker's attitudes and communication enhancers	LLOD modelling of cybersecurity terminology data		modelling an ontology of conceptual metaphors through semantic frames	
T1.2 Resources	resources related to a correlation between emotions/sentiment typesand categories of explicit and implicit offence						ontologies related to public opinion or public sentiment, parliamentary topics?	
T1.3 Interlinking	interlinking of sentiment/emotion classes to offensive language categories		Interlinking of concepts accross different languages using multilingual diachronic corpora?	multilingual aspects of discourse markers; interlinking semantics of speaker's attitudes, communication enhancers, discource relations	applying LLOD for bilingual termbase data linking			
T1.4 Sources quality								
T1.5 Under-resourced languages								

Figure 6: *WG1-WG4 matrix*

Initial contacts were also established at the Skopje meeting between WG2 and WG4 (cf. Figure 7), mainly focusing on Knowledge extraction (T2.1) and on Terminology and Knowledge Management (T2.5), areas which constitute current points of interest across practically all WG4 Use Cases. We also plan to strengthen this promising cooperation during GP3.

	Task 4.1 Linguistics		Task 4.2 Humanities & Social Sciences		Task 4.3 Technology		Task 4.4 Life Sciences	
	UC4.1.1 Media and Social Media	UC 4.1.2 Language Acquisition	UC 4.2.1 Humanities	UC 4.2.2 Social Sciences	UC 4.3.1 Cybersecurity	UC 4.3.2 FinTech	UC 4.4.1 Public Health	UC 4.4.2 Pharmacy
T2.1 Knowledge extraction			Diachronic ontology learning from text in a certain domain			?	knowledge extraction from public health reports; electronic health records and other datasets	Drug-Disease Relation Discovery and Labeling
T2.2 Machine Translation								
T2.3 Multilingual QA								
T2.4 WSD & EL			Semantic change in relation to WSD					
T2.5 Terminology & Knowledge Management			Methods for concept detection? Also ontoterminology?		Conceptual modelling of a specialised domain, categorising and linking terminological data, compilation of a termbase.	Automatic generation of finance-related terminology	extracting terminology from patient information portals/leaflets; analyzing terminology in medical reports	Health and Pharma related terminology

Figure 7: *WG2-WG4 matrix*

Although a dedicated discussion between WG4 and WG3 was not possible at the Skopje plenary meeting, preliminary contacts have already taken place and are to be intensified throughout GP3, especially given the importance of deep learning approaches (T3.2) in several Use Cases (e.g. UC4.2.1, UC4.3.1, and UC4.3.2, but also UC4.2.2).

	Task 4.1 Linguistics		Task 4.2 Humanities & Social Sciences		Task 4.3 Technology		Task 4.4 Life Sciences	
	UC4.1.1 Media and Social Media	UC 4.1.2 Language Acquisition	UC 4.2.1 Humanities	UC 4.2.2 Social Sciences	UC 4.3.1 Cybersecurity	UC 4.3.2 FinTech	UC 4.4.1 Public Health	UC 4.4.2 Pharmacy
T3.1 Big data & linguistic information								
T3.2 Deep learning and neural approaches for linguistic data			Application to detecting/representing semantic change in diachronic corpora. Explainable AI		Application of DL technologies for extraction of terms and terminological concept systems	Transformer based models for Sentiment Analysis in Finance		
T3.3 Linking structured ML language data across linguistic description levels								
T3.4 Multidimensional linguistic data								
T3.5 Education in linguistic data science								

Figure 8: *WG3-WG4 matrix*

In what concerns inter-WG cooperation, special emphasis should be given, on the one hand, to the fact that several datasets across UC include under-resourced languages, and on the other hand, to the topic of space and time representation, tackled by T1.1, T3.4, and UC 4.2.1.

3.2. Events

In GP2, Dimitar Trajanov and his team organized and hosted the NexusLinguarum 3rd plenary meeting, a hybrid event which took place at the Faculty of Computer Science and Engineering in Skopje, North Macedonia, from 29-30 September 2021. In addition, two events were organized within WG4 involving several of its members.

The first workshop on *Sentiment Analysis & Linguistic Linked Data* (SALLD-1), which took place on September 1, 2021, was co-located with LDK 2021 – 3rd Conference on Language, Data and Knowledge, in Zaragoza, Spain. Aiming to explore principles, methodologies, resources, tools, and applications combining Sentiment Analysis and Linguistic Linked Data, this workshop included an invited talk and five papers, with both an onsite and online audience. More information about this event – including the slides of the talks – is available at the official website: https://salld.org/.

In addition, UC4.1.1, dedicated to *Incivility in Media and Social Media*, organized the workshop on *Abusive language dataset annotation*, on September 28, 2021, co-located with the 3rd NexusLinguarum plenary meeting in Skopje. The more than 30 workshop participants had the opportunity to discuss the use case's new annotation guidelines. The main part of the workshop focused on hands-on live annotation of selected texts – of different lengths and offense gravity – which gave the UC coordination team further insight into various types of explicit and implicit categories of offensiveness. The outcomes of the workshop will contribute to further work on the annotation guidelines and will eventually lead to larger amounts of data collecting, as well as to a new annotation campaign. A more comprehensive description of this event can be found in a blog post recently published on the NexusLinguarum website.

3.3. Publications & STSM

Since the onset of the CA, especially in GP2, several publications have arisen from the work carried out in WG4, as follows:

- **Armaselu, F., Apostol, E-S., Khan, A.F., Liebeskind, C., McGillivray, B., Truică, C-O. and Valūnaitė Oleškevičienė, G. 2020.** HISTORIAE, History of Socio-Cultural Transformation as Linguistic Data Science. A Humanities Use Case. In Gromann, D., Sérasset, G., Declerck, T. McCrae, J.P., Gracia, J., Bosque-Gil, J., Bobillo, F. and Heinisch B. (eds.), *LDK 2021 Proceedings*. OASIcs, Volume 93, Dagstuhl Schloss: Leibniz-Zentrum für Informatik.
- **Carvalho, S. and Kernerman, I. 2021.** An overview of NexusLinguarum use cases: Current status and challenges. *Lexicala Review*, 29. https://lexicala.com/review/2021/an-overview-of-nexuslinguarum-use-cases-current-status-and-challenges/

- **Dobreva, J., Jofche, N., Jovanovik, M. and Trajanov, D. 2020.** Improving NER Performance by Applying Text Summarization on Pharmaceutical Articles. In *International Conference on ICT Innovations 2020*. Cham: Springer. 87-97.
- **Dobreva, J., Jovanovik M. and Trajanov, D. 2021.** DD-RDL: Drug-Disease Relation Discovery and Labeling. In *International Conference on ICT Innovations 2021*, Cham: Springer.
- **Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L.T. and Trajanov, D. 2020.** Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access 8* (2020): 131662-131682.
- **Rokas, A., Rackevičienė, S. and Utka, A. 2020.** Automatic Extraction of Lithuanian Cybersecurity Terms Using Deep Learning Approaches. In *Human Language Technologies – The Baltic Perspective. Proceedings of the Ninth International Conference Baltic HLT 2020*. IOS Press.

Moreover, various papers and abstracts have been submitted to peer-reviewed journals and conferences:

- **Armaselu, F., Apostol, E-S., Khan, A. F., Liebeskind, C., McGillivray, B., Truică, C.-O., Utka, A., Valūnaitė Oleškevičienė, G. and Van Erp, M**. "LL(O)D and NLP Perspectives on Semantic Change for Humanities Research" (submitted to the *Semantic Web Journal*, CFP: Special Issue on Latest Advancements in Linguistic Linked Data).
- **Fahad Khan, A., Chiarcos, C., Declerck, T., Gifu, D., González-Blanco García, E., Gracia, J., Ionov, M., Labropoulou, P., Mambrini, F., McCrae, J. P., Pagé-Perron, E., Passarotti, M., Ros Muñoz, S. and Truica, O.-C.** "When Linguistics Meets Web Technologies. Recent advances in Modelling Linguistic Linked Open Data" (submitted to the *Semantic Web Journal*, CFP: Special Issue on Latest Advancements in Linguistic Linked Data)
- **Lewandowska-Tomaszczyk, B., Žitnik, S., Bączkowska, A., Liebeskind, C., Mitrovic, J. and Valuntaite, G.** "LOD-connected offensive language Ontology and Tagset Enrichment" (submitted to the *1st Workshop on Sentiment Analysis and Linguistic Linked Data*)
- **Rackevičienė, S., Mockienė, L., and Utka, A**. "Terminology of Cyber Domain: Some Insights into the Conceptual and Linguistic Dimensions" (abstract of the presentation delivered at the international conference *Terminology - Heritage and Modernity: II International Conference*. Tbilisi - Ivane Javakhishvili Tbilisi State University. Book of abstracts).
- **Rackevičienė, S., Utka, A., Bielinskienė, A. and Rokas, A**. "Annotation of cybersecurity terminology: methodology, problems and results" (abstract of the presentation delivered at the international conference *Moksliniai, administraciniai ir edukaciniai terminologijos lygmenys = Scientific, Administrative and Educational Dimensions of Terminology: 4th international conference on terminology*, 21-22 October 2021, Vilnius. Book of abstracts).

- **Rackevičienė, S., Utka, A., Mockienė, L. and Rokas, A.** "Methodological Framework for the Development of an English-Lithuanian Cybersecurity Termbase" (submitted to *Studies about Languages* – submission accepted).
- **Rackevičienė, S., Utka, A., Bielinskienė, A. and Rokas, A.** "Distribution of Terms Across Genres in the Lithuanian Cybersecurity Corpus" (submitted to *Respectus Philologicus*).
- **Utka, A., Mockienė, L., Laurinaitis, M., Rackevičienė, S., Rokas, A. and Bielinskienė, A.** "Corpora for Bilingual Terminology Extraction in Cybersecurity Domain" (extended abstract of the presentation delivered at the *CLARIN Annual Conference 2021*, 27 – 29 September 2021 Virtual Edition. Book of abstracts).
- **Valūnaitė Oleškevičienė, G., Liebeskind, C., Trajanov, D., Silvano, P., Chiarcos, C. and Damova, M.** "Speaker Attitudes Detection through Discourse Markers Analysis" (submitted to the Workshop on *Deep Learning and Neural Approaches for Linguistic Data*. Skopje, North Macedonia & online, 30 September 2021. Book of abstracts).
- **Žitnik,S., Lewandowska-Tomaszczyk, B., Bączkowska, A., Liebeskind, C., Valuntaite, G. and Mitrovic, J.,** "Detecting Offensive Language: A New Approach to Offensive Language Data Preparation" (paper submitted to *Natural Language Engineering*).

Another relevant type of activity concerns the Short Term Scientific Missions (STSM). So far, four STSM have taken place with special focus on WG4 and its use cases, as follows:

- Feb-March 2020: Giedre Valunaite-Oleskeviciene: *Creating a multilingual corpus for formulaic language (multiword expressions) research*. Host institution: Jerusalem College of Technology. Israel. report
- Aug 2021: Liudmila Mockiene: *Corpus Analysis of Covid and health-related metaphors*. Host institution: Institute of Information and Communication Technologies at Bulgarian Academy of Sciences, Department of AI and Language. Bulgaria. report
- Aug 2021: Giedre Valunaite-Oleskeviciene: *Researching discourse markers expressing opinion with machine learning techniques in a multilingual corpus*. Host institution: Mozaika. Bulgaria. report
- Aug-Sep 2021: Kostadin Mishev: *Machine learning for detecting and interpreting language phenomena in survey data*. Host institution: Mozaika. Bulgaria. report

Concluding remarks and next steps

This intermediate report outlined the progress within WG4 and its Use Cases, especially in the last few months of GP2 (May – October 2021), given that a comprehensive state-of-the-art had already been provided in the first deliverable, in April 2021. Overall, extensive work has been carried out, with several UC stabilizing their conceptualization stages and beginning implementation. New datasets and tools are being developed, with a strong multilingual focus. Intra-WG cooperation has been fostered and supported by the complementary linguistic and computational backgrounds of our members, especially in what concerns devising LD-compliant solutions for resource creation and dissemination. The topics of Sentiment Analysis and LLD, as well as knowledge organization and extraction, have clearly emerged as relevant across UC, and will be further explored throughout the CA.

On the other hand, inter-WG cooperation has been successfully unfolding as well. Stronger connections have been established with WG1 in the first two Grant Periods, mainly addressing the initial/current Use Cases' challenges regarding modelling and interlinking, but it is expected that the engagement with other tasks within WG1 becomes increasingly visible. Contacts with WG2 will become more regular during GP3, especially given the importance of terminological data (not only their extraction, but also their representation) across the various UC. Regarding collaboration with WG3, although deep learning seems to be the most prominent topic at the moment, multidimensional LD is gaining traction, particularly the challenges underlying the modelling of such data across space and time, an issue which will start being discussed in more depth in GP3 within WG1, WG3 and WG4 (especially through UC 4.2.1) joint sessions.

To help foster the aforementioned intra- and inter-WG cooperation, WG4 aims to further encourage UC-related applications for STSM or Virtual Mobility grants, as well as the collaboration with other networks and projects (such as CLARIN, the European Language Grid, among others). In addition, the WG members plan to organize two events in GP3: a workshop proposal for the second edition of SALLD has been submitted to LREC 2022; and an international conference on *Explicit and Implicit Abuse* is being prepared for the Spring of 2022, in Jerusalem, within the scope of UC4.1.1. Finally, another call for new Use Cases is to be launched in GP3, aiming to get contributions from areas which may be relevant for NL but are not yet covered in the current WG4 landscape.

References

Bentz, C., Ruzsics, T., Koplenig, A. and Samardžić, T. 2016. A Comparison Between Morphological Complexity Measures: Typological Data vs. Language Corpora. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity* (CL4LC), pp. 142-153.

International Standardization Organization. 2016. *ISO 24617-8:2016 Language resource management – Semantic annotation framework (SemAF) – Part 8: Semantic relations in discourse, core annotation schema (DR-core)*. Geneva: ISO.

MacWhinney, B. 2000. *The CHILDES Project: Tools for analyzing talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Štrkalj Despot, K. and Ostroški Anić, A. 2021. A War on War Metaphor: Metaphorical Framings in Croatian Discourse on Covid-19. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 47(1), doi.org/10.31724/rihjj.47.1.6.