Nexus Linguarum

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Deliverable 4.3

Final Activity Report (months 25-54). Working Group 4: *Use Cases and Applications*

Main authors: Sara Carvalho, Ilan Kernerman

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

25 April 2024

Project Acronym	NexusLinguarum
Project Title	European network for Web-centred linguistic data science
COST Action	18209
Starting Date	26 October 2019
Duration	54 months
Project Website	https://nexuslinguarum.eu/
Chair	Jorge Gracia
Main authors	Sara Carvalho and Ilan Kernerman
Contributors	Florentina Armaselu, Mariana Damova, Kristina Despot, Dagmar Gromann, Gordana Hržica, Mietta Lennes, Barbara Lewandowska-Tomaszczyk, Barbara McGillivray, Maciej Ogrodniczuk, Petya Osenova, Ana Ostroški Anić, Sigita Rackevičienė, Marko Robnik-Šikonja, Jouni Tuominen, Dimitar Trajanov, Slavko Žitnik
Reviewers	NexusLinguarum core group team
Version | Status	05 | final version
Date	25 April 2024

Acronym 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
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
VMG	Virtual Mobility Grant
WG	Working Group

Table of Contents

Executive Summary

This report summarises the evolution of the tasks and use cases explored in Working Group 4 (WG4) of the NexusLinguarum COST Action (CA 18209) from November 2021 to April 2024. It builds upon the first and second deliverables (D4.1 – Use Case Description and Requirements Elicitation and D4.2 – Intermediate Activity Report), submitted in April and October 2021, respectively, as well as a journal article published in July 2021 (*Lexicala Review*, 29: 26-72), which comprise a comprehensive description of the use cases and the roadmap and intermediate milestones regarding their implementation. In addition to the progress of each of the four tasks and, in particular, of the nine use cases, this conclusive report elicits the solid inter-WG cooperation, as well as WG4's achievements.

1. Introduction

NexusLinguarum's Working Group 4 explores use cases and applications in which the Action's relevant methodologies, technologies, and standards have been tested and validated. Since D4.1 in April 2021, the number of members has continued to grow to a total of 138, and the members' multidisciplinary backgrounds have constituted one of WG4's major assets. During the entire period of the CA, 23 broad WG bi-monthly meetings were held, alongside 31 meetings of the core team and numerous meetings of the WG4 leaders.

The structure underpinning WG4 incorporates this interdisciplinary approach and benefits from having dual leaders per Task, one expert in Linguistics and one in Computer Science. Given the wide range of selected domains (Linguistics, Humanities and Social Sciences, Technology, and Life Sciences), a second level was formed to accommodate the actual use cases and applications. The structure is depicted in Figure 1, with each of the four Tasks incorporating its Use Cases (UCs).

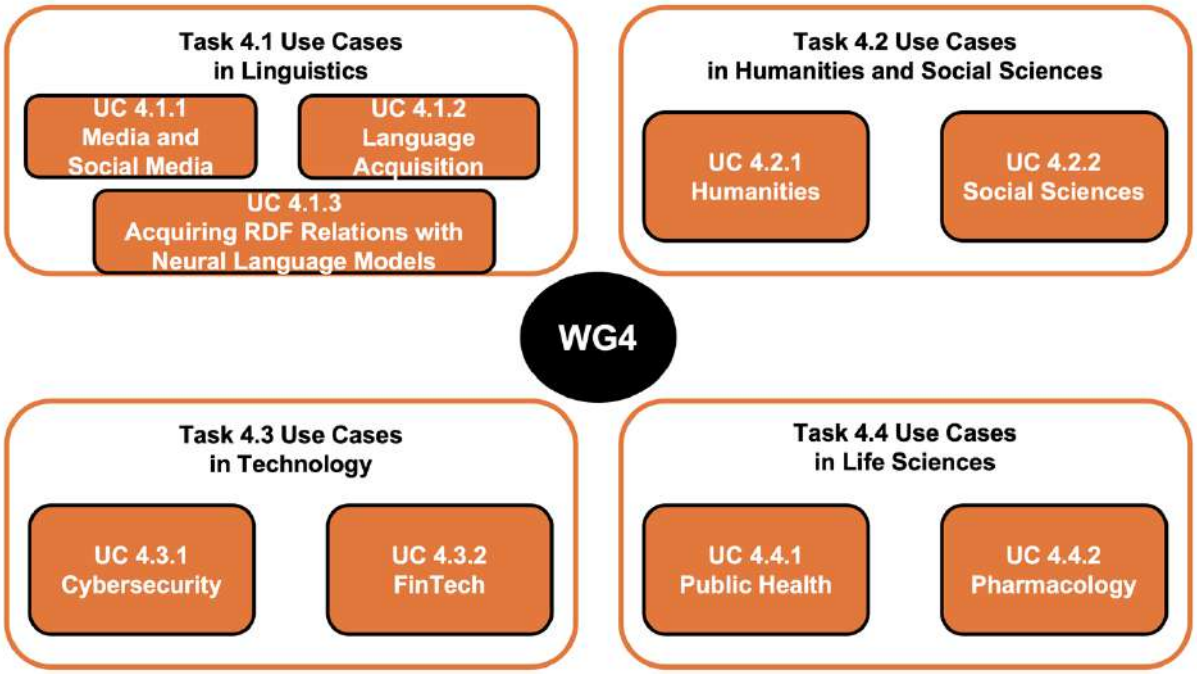


Figure 1. WG4's structure of Tasks and Use Cases

The WG's core team has remained mostly solid throughout the CA's lifetime, with minor changes, 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
UC4.1.3 coordinator	Dagmar Gromann	Austria
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 (until December 2021) Tilia Ellendorff (from January 2022)	Romania Switzerland
T4.3 leader – computational	*Valentina Janev (until October 2021) Dimitar Trajanov (from November 2021)	Serbia North Macedonia
UC 4.3.1 coordinator	Sigita Rackevičienė	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 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 (D4.1, 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 intensified during GP2, as reported in the second deliverable (D4.2, submitted in October 2021), and continued to expand substantially in the subsequent GPs (November 2021 – April 2024). Close collaboration both within and outside of NexusLinguarum also formed a critical part of WG4's mission. Moreover, the work developed in WG4 fostered a wide range of exchange and dissemination activities, as outlined in Section 3.

2. Task and Use Case updates

2.1 Task 4.1 Use Cases in Linguistics

Task Leaders

- Kristina Š. Despot, Institute of Croatian Language and Linguistics (linguistics)
- Slavko Žitnik, University of Ljubljana (computational)

Use Cases

- UC 4.1.1: Use Case in Incivility in Media and Social Media
- UC 4.1.2: Use Case in Language Acquisition
- UC 4.1.3: Use Case in Acquiring RDF Relations with Neural Language Models

Overview

The task investigates how linguistic data science and a richer understanding of language, based on the techniques explored in WG3, can benefit research in linguistics (e.g., lexicography, terminology, typology, syntax, comparative linguistics, computational linguistics, corpus linguistics, phonology, etc.). General tasks within this endeavour include: staying up-to-date with the state-of-the-art (SOTA) in the usage of Linked Open Data (LOD) in Linguistics; creating a document that describes requirements elicitation and use case definition (M18); submitting intermediate and final activity reports (M24 and M48); and producing scientific papers on the in-use applications of Linked Linguistic Open Data (LLOD), Natural Language Processing (NLP), and linguistic big data (M48).

More specific tasks were accomplished within particular use cases that are described in detail. The task consists of three use cases focusing on Media and Social Media (UC4.1.1), Language Acquisition (UC4.1.2), and Acquisition of RDF Relations with Neural Language Models (UC4.1.3). The first deals with offensive language analysis from a linguistic perspective, applying text analytics tools to support linguistic phenomena. The second focuses on the development of tools for language acquisition and analysis. The third addresses the use of pre-trained neural language models for the acquisition of RDF relations across natural languages.

The three use cases operated in parallel, employing focus groups for specific tasks. They were very active and productive, producing relevant results in the form of lectures and workshops on specific topics, conference presentations, specific targeted joint tasks, and the publication of datasets, tools, and papers. Both task leaders were actively involved in these use cases and also organised overall meetings where new ideas, collaborations, and possible future directions were discussed.

In conclusion, the task has demonstrated the practical applications of linguistic data science and the valuable insights gained from the exploration of advanced LLOD techniques and the collaboration between linguists and computer scientists. The accomplishments within the specific use cases highlight the task's contribution to offensive language detection, analysis and language acquisition tools, and acquisition of RDF relations via neural language models.

2.1.1 UC4.1.1 Use Case in Incivility in Media and Social Media

Coordinator: Barbara Lewandowska-Tomaszczyk (University of Applied Sciences in Konin)

Overview

The principal aim of this use case was to develop cumulative knowledge on the identification and extraction of incivility of media discourse content in online newspaper texts and social media, as well as to conduct a systematic survey of ways to create an infrastructure regarding offensive language data sharing. The UC team modified existing offensive language taxonomies and proposed a Simplified Taxonomy of Offensive Language (SOLT). Following two series of annotation practices, the SOLTs were implemented in Czech, English, Hebrew, Lithuanian and Polish, and exemplified and analysed in Croatian, Slovene and Kazakh. The analyses cover both implicitly and explicitly abusive content in (i) intentionally offensive messages (explicit and implicit), (ii) hate speech, (iii) personal insults, and (iv) abusive words or phrases (vulgarisms) in jokes and in cursing, etc. Researched materials include online newspaper articles and comments, online posts, forum audiences as well as public posts of one-to-one, one-to-many, many-to-one and many-to-many types. The gold standard of the best examples with their annotations has been included as an LLOD schema.

Resources, methods, tools/technologies, languages used

We have used exploratory techniques leveraging pure linguistic knowledge (i.e. theory, SketchEngine corpora and tools) and computational linguistic processing of available data (i.e. (non-)contextual embeddings and BERT). As a result, we proposed a taxonomy of offensive language, consisting of:

- Main offensive language categories: OFFENSIVE was selected as the umbrella term being the most general category, while other labels were organized into a three-level hierarchy, depending on the particular level of specificity
- implicit or explicit *expressiveness*
- group, individual or mixed *target type*
- internal, external and border *target level*
- A*spects* in which an offense can be represented
- an indicator of whether the language is *figurative* or not
- *Categories* of implicit language of offense

Some of the **resources** used include:

- Big data: national language corpora, media and social media repositories, platforms; CLARIN https://lindat.mff.cuni.cz/services/kontext/corpora/corplist, eval-data
- Hate speech datasets: hatespeechdata.com (Derczynski & Vidgen, 2020)
- Samplers: small corpora of social media such as NLTK (Natural Language Toolkit), a small collection of web texts, parts of EUROPARL
- Datasets of languages represented in the use case.

As regards the **methods**, we highlight:

- Data identification and acquisition – Media Studies and Corpus Linguistics
- Modelling Hate-Event (HE) structure (Lexical approaches, Prototypical Event Semantics, Cognitive Corpus Linguistics)
- Incivility/abuse identification scales (explicit, implicit) – Statistical and qualitative approaches
- Abusive language tagset annotation identification and surveys
- Computational methodologies (cf **Tools**)
- Enrichment/formal simplification of explicit and implicit language tagsets towards offensive language extraction

Some of the **tools/technologies** used include:

Text categorization: Naive Bayes, Support Vector, Machine and Logistic Regression. Open-source implementations.

The traditional methods (Naive Bayes, Support Vector Machines, Logistic Regression) have been useful for explicit abusive language, while contextual Deep learning models based (e.g., ELMo) on transformer architectures, such as BERT, GPT, ELECTRA, have been tested for the more complex tasks.

Semantically-based identification of Multi-Word Expressions: Spyns & Odijk (eds.), 2013 - Equivalence Class Method (ECM)

Classificatory hate speech models: Davidson et al. (2017), FastText, Neural Ensemble, BERT

NLP extraction tools: Keyword-based approaches SemEval 2019 e.g., http://alt.qcri.org/semeval2019/index.php?id=tasks; Naive Bayes, Support Vector Machine and Logistic Regression; Multiple-view Stacked Support Vector Machine (mSVM) – multiple classifiers application

Semantic Annotation Platforms: INCEpTION annotation tools (inception-project.github.io)

Languages: Croatian, Czech, English, Hebrew, Kazakh, Lithuanian, Polish

Tasks

The following tasks were conducted in parallel throughout the Action lifetime.

T1. Description of details of the use case objectives and implementation: (abuse, implicitness/explicitness, emotions/sentiments, hate speech); selecting languages

T2. Selection of English hate speech datasets for analysis

T3. Survey of accessible sets of abuse language dimensions

T4. Identification of explicit vs implicit abuse identificatory and classification criteria – direct literal vs indirect and figurative

T5. Development of abusive language identification scales

T5.a Manual tagging of the selected data based on new decisions and scales in English and other languages (all members) – Task use workshop devoted to this activity at one of the stationary WG4 meetings)

T6. Survey of automatic annotation tools and implementation of the baseline model (Annotation I.)

T7. Abusive language tagset enrichment/formal simplification proposals: Annotation II.

T8. Survey of LLOD infrastructure relevant to the task topic (ontologies of opinion mining, etc.). Infrastructure proposals of abusive hate speech data sharing. LLOD offensive language schema publication.

Roadmap

Year 1: Survey/selection of corpora

- Development of offensive language taxonomy models I and 2 (extended)– annotation campaign I

Year 2: Publications/workshops on results of Taxonomy 2.

Year 3: Simplified offensive language taxonomy SOLT and annotation II on 5 Languages, Results of annotation II.

Final phase: Extension of the SOLT to other languages and development of the LLOD schema for offensive language annotation and retrieval.

Four workshops were held to discuss the computational aspects involved in each of the planned tasks (2020, 2022, 2023 (2)):

- Development of incivility/abuse identification scales (explicit, implicit)
- Identification of tagset annotation tools
- Application and modification of the taxonomy and tagset tools to Croatian, Czech, English, Hebrew, Kazakh, Lithuanian, Polish.
- Development of the LLOD schema for Offensive Language Identification.

Strategy

- The main aim of this use case was to build cumulative knowledge on the identification and extraction of incivility of media discourse content in online newspaper texts and social media;
- The first stage toward the main objective was the identification of abusive language corpora and their annotation and extraction tools;
- The main strategy covered the development of richer abuse event identification structures and identification scales, together with scrutinizing particular steps in the annotation campaigns.
- The main outcome involves the development of relevant offensive language scales and their respective use for LLOD.

Modifications of the original taxonomy, evident in the numerous publications based on the NexusLinguarum works of the UC and used eventually in the last annotation cycle, are finally represented in seven languages (Croatian, Czech, English, Hebrew, Kazakh, Lithuanian, Polish). The outcome, as planned in the Milestones, is a development of the more universal LLOD taxonomy of offensive languages available online and presented at the last NexusLinguarum general meeting in Athens, in 2024.

Findings

The outcomes of this UC fully conform to all planned activities envisaged in the first year of the Action lifetime and contribute to the development of the working LLOD criteria, taxonomy and identification of the offensive language in media texts. The UC4.1.1 team prepared a clear, multi-level and multilingual taxonomy of offensive language, scrutinizing semantic relationships between words on the basis of linguistic knowledge and the results of research from various corpus tools. The results are likely to have a more universal application to other languages.

Deliverables

(1) Publications

- Arhar Holdt, Špela, Simon Krek, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Jaka Čibej, Eva Pori, Luka Terčon, Tina Munda, Slavko Žitnik, Nejc Robida, Neli Blagus, Sara Može, Nina Ledinek, Nanika Holz, Katja Zupan, Taja Kuzman, Teja Zupančič, Iza Škrjanec, Dafne Marko, Lucija Jezeršek, Ana Marija Zajc, Training corpus SUK 1.0, Ljubljana: Institut Jožef Stefan, 2022, CLARIN.SI data & tools, ISSN 2820-4042, http://hdl.handle.net/11356/1747.
- Arhar Holdt, Špela, David Bordon, Jaka Čibej, Kaja Dobrovoljc, Polona Gantar, Jakob Lenardič, Tina Munda, Eva Pori, Nejc Robida, Luka Terčon, Slavko Žitnik, Slovenski učni korpus: množici SUK 1.0 in Janes-Tag 3.0 : poročilo projekta Razvoj slovenščine v digitalnem okolju : aktivnost DS1.2, Ljubljana: Univerza v Ljubljani, Center za jezikovne vire in tehnologije, 2023.
- Bączkowska, Anna (2021). "You're too thick to change the station" – Impoliteness, insults and responses to insults on Twitter. Topics in Linguistics 22(2), 62-84.
- Bączkowska, Anna (2022). Explicit and implicit offensiveness in dialogical film discourse in Bridgit Jones films. International Review of Pragmatics 14(2). 198–225.
- Bączkowska, Anna (2022). "Hope you have a shit birthday you fat cunt" – cognitive strategies, rhetorical figures and linguistic means used in insulting Tweets. Forum Filologiczne Ateneum 1(10), 9-25.
- Bączkowska, Anna (2023). "Implicit offensiveness from linguistic and computational perspectives: A study of irony and sarcasm", Lodz Papers in Linguistics, 19(2), 353-383.
- 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 49(1), 79-97.
- Bolatbek, M. and Sh. Mussiraliyeva (2023). Detection of extremist messages in web resources in the Kazakh language. Lodz Papers in Pragmatics 19 (2). 415-425.
- Chiarcos, Christian, Purificação Silvano, Mariana Damova , Giedre Valunaite Oleškeviciene , Chaya Liebeskind , Dimitar Trajanov , Ciprian-Octavian Truică, Elena-Simona Apostol and Anna Bączkowska (2023). Building an OWL-Ontology for Representing, Linking and Querying SemAF Discourse Annotations. Rasprave 49(1), 117-136.
- Damova, Mariana, Kostadin Mishev, Giedrė Valūnaitė-Oleškevičienė, Chaya Liebeskind, Purificação Silvano, Dimitar Trajanov, Ciprian-Octavian Truica, Elena-Simona Apostol, Christian Chiarcos, and Anna Baczkowska. 2023. Validation of Language Agnostic Models for Discourse Marker Detection. In Proceedings of the 4th Conference on Language, Data and Knowledge, pages 434–439, Vienna, Austria. NOVA CLUNL, Portugal.
- Despot, Kristina Š. and Veale, Tony: "Somewhere along your pedigree, a bitch got over the wall!" – A data-driven approach to a typology of implicitly offensive language. Researching and Applying Metaphor. Madrid, Alcala de Henares, Spain. 27 – 30 June 2023. Madrid, Alcala de Henafres, Spain 27-30 June 2023.

- Despot, Kristina Š., Brač, Ivana and Filipić, Lobel. The role of metaphors in shaping public discourse around social media. CILID – 2nd Conference on Digital Linguistics. Alicante, Spain. 3 – 5 May 2023.
- Despot, Kristina, Ostroški Anić, Ana and Veale, Tony. Vanity Degrees, Parasitic Non-Jobs, and Geniuses by Donkey Standards – A Typology of Implicitly Offensive Language. CLARC 2023: Language and Linguistic Data. 28 – 30 September 2023. Rijeka.
- Dontcheva-Navrátilová, O., & Povolná, R. (2023). Czech Offensive Language: Testing a Simplified Offensive Language Taxonomy. In Cavalho, S., Gracia, J., Khan, A. H., McCrae, J., Ostroški, A., Fromann, D., Spahiu, B., Heinisch, B., Salgado, A. (Eds.). LDK2023 - Proceedings of the 4th Conference on Language, Data and Knowledge, pp. 634-639.
- Dontcheva-Navrátilová, O., & Povolná, R. (2023). Offensive language in media discussion forums: A pragmatic analysis. Lodz Papers in Pragmatics, 19(2), pp. 223-238. https://dx.doi.org/10.1515/lpp-2023-0012
- Dontcheva-Navrátilová, O., & Povolná, R. (2023). Czech Offensive Language: Testing a Simplified Offensive Language Taxonomy (conference presentation). 4th Conference on Language, Data and Knowledge, SALLD-3 Workshop.
- Gec, Saandi, Vlado Stankovski, Marko Bajec, Slavko Žitnik, "Simulacija in izboljšava prometnih tokov : primer na dveh izbranih slovenskih križiščih", Uporabna informatika, 2022, vol. 30, no. 3, pp. 151-168, DOI: 10.31449/upinf.170.
- Kern Pipan, Matej, Ana Rosa Guzmán Carbonell, Ioannis Konstantinidis, Dejan Lavbič, Slavko Žitnik, Miha Jesenko, Thashmee Karunaratne, Multilingual ontology repository and user interface : DE4A semantic interoperability : SEMIC Conference 2022, Data spaces in an interoperable Europe, 6th of December 2022, Brussels.
- Klamra Cezary, Wojdyga Grzegorz, Żurowski Sebastian, Rosalska Paulina, Kozłowska Matylda, Ogrodniczuk Maciej, “Devulgarization of Polish texts using pre-trained language models”. In: Derek Groen, Clélia de Mulatier, Maciej Paszynski, Valeria V. Krzhizhanovskaya, Jack J. Dongarra, Peter M. A. Sloot (editors) Computational Science – ICCS 2022, Lecture Notes in Computer Science vol. 13351, pp. 49–55. Springer International Publishing, Cham, 2022. https://doi.org/10.1007/978-3-031-08754-7_7
- Klemen, Matej and Slavko Žitnik, "Neural coreference resolution for Slovene language", Computer science and information systems, 2022, vol. 19, iss. 2, pp. 495-521, DOI: 10.2298/CSIS201120060K.
- Klemen, Matej, Aleš Žagar, Timotej Knez, Frenk Dragar, Marko Robnik Šikonja, Slavko Žitnik, Magdalena Gapsa, Mojca Brglez, Katarina Aleksandra Brezovar, Špela Vintar, "34. evropska povola šola logike, jezika in informatike ESSLLI 2023", Slovenščina 2.0 : empirične, aplikativne in interdisciplinarne raziskave, 2023, vol. 11, no. 2, pp. 84-91, DOI: 10.4312/slo2.0.2023.2.84-91.
- Kljun, Maša, Matija Teršek, Slavko Žitnik, "Pomenska analiza kategorij sovražnega govora v obstoječih označenih korpusih", Uporabna informatika, 2022, vol. 30, no. 1, pp. 3-18, DOI: 10.31449/upinf.151.
- Knez, Timotej, Marko Bajec, Slavko Žitnik, "ANGLEr : a next-generation natural language exploratory framework", In: Research challenges in information science : 16th International Conference, RCIS 2022, Barcelona, Spain, May 17-20, 2022 : proceedings, Renata Guizzardi (ed.), Jolita Ralyté (ed.), Xavier Franch (ed.), Cham: Springer, cop. 2022, pp. 761-768, DOI: 10.1007/978-3-031-05760-1_53.

- Knez, Timotej, Domen Gašperlin, Marko Bajec, Slavko Žitnik, "Blockchain-based transaction manager for ontology databases", Informatica, [Print ed.], 2022, vol. 33, no. 2, pp. 343-364, DOI: 10.15388/22-INFOR490.
- Knez, Timotej, Slavko Žitnik, "Temporal relation extraction from clinical texts using knowledge graphs", In: Research challenges in information science : information science and the connected world : 17th International Conference, RCIS 2023, Corfu, Greece, May 23–26, 2023 : proceedings, Selmin Nurcan (ed.), Cham: Springer, cop. 2023, pp. 493-500, DOI: 10.1007/978-3-031-33080-3_30.
- Knez,Timotej, Slavko Žitnik, "Word in context task for the Slovene language", In: Language, data and knowledge 2023: LDK 2023 : proceedings of the 4th Conference on Language, Data and Knowledge : 12-15 September 2023, Vienna, Auppia, Sara Carvalho (ed.), [S. l.]: Nova Clunl, 2023, pp. 322-327.
- Knez, Timotej, Slavko Žitnik, "Event-centric temporal knowledge graph conppuction : a survey", Mathematics, Dec. 2023, vol. 11, iss. 23, [article no.] 4852, pp. 1-32, DOI: 10.3390/math11234852.
- Lewandowska-Tomaszczyk, B. (2022). A simplified taxonomy of offensive language (SOL) for computational applications. Konin Language Studies 10(3). 213–227.
- Lewandowska-Tomaszczyk, B., Slavko Žitnik, Anna Bączkowska, Chaya Liebeskind, Jelena Mitrović, Giedrė Valūnaitė Oleškevičienė (2021). LOD-connected offensive language ontology and tagset enrichment. CEUR Workshop Proceedings, 135-150.
- Lewandowska-Tomaszczyk, B., Slavko Žitnik, Chaya Liebeskind, Giedre Valunaite Oleskevicienė, Anna Bączkowska, Paul A. Wilson, Marcin Trojszczak, Ivana Brač, Lobel Filipić, Ana Ostroški Anić, Olga Dontcheva-Navratilova, Agnieszka Borowiak, Kristina Despot, Jelena Mitrović (2023). Annotation Scheme and Evaluation: The Case of Offensive Language. Rasprave : Časopis Instituta za hrvatski jezik i jezikoslovlje , Vol. 49 No. 1. 155-175.
- Lewandowska-Tomaszczyk, B., Anna Bączkowska, Chaya Liebeskind, Giedre Valunaite Oleskevicienė, Slavko Žitnik. (2023). An integrated explicit and implicit offensive language taxonomy. Lodz Papers in Pragmatics, 19(1), 7-48.
- Litvak, M., Natalia Vanetik, Chaya Liebeskind, Omar Hmdia and Rizek Abu Madeghem Offensive language detection in Hebrew: can other languages help?.in Proceedings of the Language Resources and Evaluation Conference, June 2022, Marseille, France, European Language Resources Association, p. 3715-3723
- Liebeskind, C., Vanetik, N., & Litvak, M. (2023). Hebrew offensive language taxonomy and dataset. Lodz Papers in Pragmatics, 19(2), 325-351.
- Litvak, M., Vanetik, N., Liebeskind, C., Hmdia, O., & Madeghem, R. A. (2022, June). Offensive language detection in Hebrew: can other languages help?. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 3715-3723).
- Liebeskind, C., Liebeskind, S., & Yechezkely, S. (2021, October). An Analysis of Interaction and Engagement in YouTube Live Streaming Chat. In 2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI) (pp. 272-279). IEEE.

- Luzzon, E., & Liebeskind, C. (2023, July). JCT_DM at SemEval-2023 Task 10: Detection of Online Sexism: from Classical Models to Transformers. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023) (pp. 739-743).
- Možina, Marko, Slavko Žitnik, Barbara Koroušić-Seljak, Tome Eftimov, "Enhancing food composition databases : predicting missing values via knowledge graph embeddings", In: KDD2023 : 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining: Long Beach, August 6-10 [2023], [New York]: Association for Computing Machinery, 2023, pp. 1-6.
- Natanya, T., & Liebeskind, C. (2023). Clickbait detection in Hebrew. Lodz Papers in Pragmatics, 19(2), 427-446.
- Ostroški Anić, Ana; Despot, Kristina Š; Terčon, Luka. Detecting patterns of implicit offensive language in multilingual data (poster). First UniDive Workshop and 1st General Meeting. 15-16 March 2023. Paris-Saclay University, France. First UniDive Workshop and 1st General Meeting. 15-16 March 2023. Paris-Saclay University, France.
- Prelevikj, P. and Slavko Žitnik, "Multilingual named entity recognition and matching using BERT and dedupe for Slavic languages", In: Proceedings of 8th BSNLP Workshop on Balto-Slavic Natural Language Processing, co-located with the 16th European Chapter of the Association for Computational Linguistics (EACL), April 20, 2021, Bogdan Babych (ed.), Senja Pollak (ed.), ppoudsburg: The Association for Computational Linguistics, 2021, pp. 80-85, tabele, ISBN 978-1-954085-14-5, https://www.aclweb.org/anthology/2021.bsnlp-1.9/.
- Šoltes, Tjaša, Marko Bajec, Iztok Lebar Bajec, Kaja Gantar, Slavko Žitnik, "Online-notes system: real-time speech recognition and translation of lectures", In: Research challenges in information science : information science and the connected world : 17th International Conference, RCIS 2023, Corfu, Greece, May 23–26, 2023 : proceedings, Selmin Nurcan (ed.), Cham: Springer, cop. 2023, pp. 485-492, DOI: 10.1007/978-3-031-33080-3_29.
- Štrkalj Despot, Kristina; Ostroški Anić, Ana; Veale, Tony. 2023. “Somewhere along your pedigree, a bitch got over the wall!” A proposal of implicitly offensive language typology. Lodz Papers in Pragmatics 19/2. 385–414. doi: 10.1515/lpp-2023-0019.
- Valūnaitė-Oleškevičienė, Giedrė, Selmistraitis, Linas, Utka, Andrius and Gudelis, Dangis. "Offensive language in user-generated comments in Lithuanian" Lodz Papers in Pragmatics, vol. 19, no. 2, 2023, pp. 239-254. https://doi.org/10.1515/lpp-2023-0013
- Žitnik, Slavko and Frenk Dragan, SloBENCH evaluation framework, [S. l.]: CLARIN.SI, 2021, CLARIN.SI data & tools, http://hdl.handle.net/11356/1469.
- Žitnik, Slavko, Karmen Kern Pipan, Miha Jesenko, Dejan Lavbič, "Semantic reusable web components: a use case in e-government interoperability", Uporabna informatika, 2022, vol. 30, no. 4, pp. 256-269, DOI: 10.31449/upinf.189.
- Žitnik, Slavko, Neli Blagus, Marko Bajec, "Target-level sentiment analysis for news articles", Knowledge-based systems, Aug. 2022, vol. 249, pp. 1-14, DOI: 10.1016/j.knosys.2022.108939.

- Žitnik, Slavko, Glenn Gordon Smith, "Automated analysis of postings in fourth grade online discussions to help teachers keep students on-track", Interactive learning environments, 2023, pp. 1-26, DOI: 10.1080/10494820.2023.2204327.
- Žitnik, S., Lewandowska-Tomaszczyk, B., Bączkowska A., Liebeskind, C., Valūnaitė Oleškevičienė, G., and Mitrović, J. (2023). Detecting Offensive Language: A New Approach for Offensive Language Data Preparation. The Second International Israel Data Science Initiative Conference. Dead Sea, Israel (Poster).

(2) Conference/workshop presentations

- Bączkowska, Anna, Lewandowska-Tomaszczyk, Barbara, Valunaite Oleskevicienė, Giedrė. **Annotation Taxonomy of Implicit Offensive Language**. Workshop *Taxonomy and annotation of offensive language: Implicitness*. *Nexus workshop days in Jerusalem*. Jerusalem College of Technology, Jerusalem, May 23-24 2022.
- Bączkowska, Anna. **Hateful language on Polish Twitter: availability, potential and problems**. *International Conference Contacts & Contrasts. Languages in Cultural Perspectives: Practices, Discourses, Cognition*. Konin, Poland, 27-29 March 2023.
- Despot, Kristina; Ostroški Anić, Ana. **Overview of Approaches to Implicitness**. Workshop *Taxonomy and annotation of offensive language: Implicitness*. *Nexus workshop days in Jerusalem*. Jerusalem College of Technology, Jerusalem, May 23–24 2022.
- Lewandowska-Tomaszczyk, Barbara, Chaya Liebeskind, Marcin Trojszczak, Slavko Žitnik, Anna Baczkowska & Giedre Valuntaite Oleškevičiene. Presenters: Chaya Liebeskind and Marcin Trojszczak. **An offensive language taxonomy and a web corpus discourse analysis for automatic offensive language identification**. *ADDA3 - Approaches to Digital Discourse Analysis*, Florida, St. Petersburg, 13-15 May 2022.
- Lewandowska-Tomaszczyk, Barbara, Slavko Žitnik, Giedrė Valūnaitė Oleškevičienė, Anna Bączkowska, Chaya Liebeskind, Marcin Trojszczak, Maria da Purificacao Moura Silvano (Presenter: Giedrė Valūnaitė Oleškevičienė). **An Integrated Explicit and Implicit Offensive Language Taxonomy**. *International Conference on Online Hate Speech*. University of Minho, Portugal, 7-9 July 2022.
- Lewandowska-Tomaszczyk, Barbara, Anna Bączkowska, Olga Dontcheva-Navrátilová, Chaya Liebeskind, Giedrė Valūnaitė-Oleškevičienė, Slavko Žitnik, Marcin Trojszczak, Renata Povolná, Linas Selmistraitis, Andrius Utka and Dangis Gudelis. **LLOD schema for offensive language simplified taxonomy (SOL) in multilingual detection and applications**. *SALLD-3 (co-located with LDK 2023)*, Vienna, Austria, 13-15 September 2023.
- Ostroški Anić, Ana; Despot, Kristina. **Challenges of detecting online hate speech: the linguistic perspective**. *Law, Society and Human Rights in the Digital Age*, University of Rijeka, Faculty of Law, 27 April 2023.
- Silvano, Purificação, Mariana Damova, Giedrė Valūnaitė Oleškevičienė, Chaya Liebeskind, Christian Chiarcos, Dimitar Trajanov, Ciprian-Octavian Truică, Elena-Simona Apostol and Anna Baczkowska. **ISO based Discourse Markers Annotated Multilingual Corpus**. *LREC 2022*, Marseille, France, 20-25 May 2022.

- Valūnaitė-Oleškevičienė, Giedrė, Linas Selmistraitis, Andrius Utka, Dangis Gudelis. **Detecting Offensive Language in Lithuanian**. *International Conference Contacts & Contrasts. Languages in Cultural Perspectives: Practices, Discourses, Cognition*. Konin, Poland, 27-29 March 2023.

(3) Other presentations

- Bączkowska, Anna. *New Approaches to Implicitness - a model for offensive language*. Presentation at the 4.1.1. Use case meeting, 16 December 2021.
- Despot, Kristina; Ostroški Anić, Ana. *Implicit Offensive Language*. Presentation at the 4.1.1. Use case meeting, 20 January 2022.
- Barbara Lewandowska-Tomaszczyk, Slavko Žitnik, Anna Bączkowska, Giedrė Valūnaitė Oleškevičienė, Chaya Liebeskind, Kristina Despot, Ana Ostroški Anić. *Implicit offensive language annotation: a new proposal.* **Implicitness in Offensive Language Workshop.** Co-located with the NexusLinguarum Workshop Days in Jerusalem (May 23-24 2022).
- Anna, Bączkowska, Barbara Lewandowska-Tomaszczyk, Slavko Žitnik, Giedrė Valūnaitė Oleškevičienė, Chaya Liebeskind, Kristina Despot, Ana Ostroški Anić. *Offensive Language Categorization for the LLOD Schema*. Presentation at the 6th Nexus Plenary Meeting, Athens, 20 March 2024.

Work in progress

A publication in preparation: *LLOD Offensive Language taxonomy: Gold standard in best examples.*

2.1.2 UC4.1.2 Use Case on Language Acquisition

Coordinator: Gordana Hržica (University of Zagreb)

Overview

A language sample consists of written or spoken text that can be used to assess the language skills of an individual speaker by speech and language pathologists for first language acquisition as well as by second language teachers. Measures for analysis have been developed and validated, but mainly for English. Web services have been developed and are mostly used for English, while very few, such as Coh-Metrix, have been adapted to other languages (Spanish, Portuguese, German). Our goal was to make use of existing language technologies to develop tools for other languages, to introduce and validate new measures for advanced analysis, and to promote the usage of speech-sample analysis in different fields, such as regular education.

In the framework of this use case, we have developed one web-based application for the analysis of language samples in Croatia. We are also currently discussing with our Lithuanian colleagues the possibility of adapting the application for Lithuania. We published four papers, presented our work eight times at scientific conferences, and supported one VMG. We had 11 group meetings online and met in smaller groups 18 times to discuss the data collection, development of new measures, and work on the web-based application.

Resources, methods, tools/technologies, languages used

A) To develop new measures for the language sample analysis we compiled a database with language samples from different languages (using the TalkBank database CHILDES and our personal networks). A new type of analysis, with automated measures of morphological richness, was applied to language samples.
B) We used the existing TalkBank database CHILDES corpora of Croatian to perform the analyses on different levels of discourse analysis.
C) We developed a new web-based application to analyse language samples in Croatian, focusing mostly on spoken language samples of children. The modules include a morphological tagger, syntactic parser, syntactic tagger, and automatic coding of repetitions.
D) We collected the data and provided transcripts of speech to enable comparisons between individual analysis in the web-based application and larger groups of age-matched individuals.

Roadmap, workflow and milestones

T1. Researching available language sample tools

During the second year of the project, we researched available tools for language sample analysis in different languages. Additionally, we explored which measures were found to be the most common and for which there is information about their validity. We especially focused on measures of lexical diversity and syntactic complexity. A gap was discovered: there are hardly any measures that focus on the morphological diversity of language samples.

T2. Researching available language technologies for participating languages

During the fourth year of the project, as part of exploring how to develop the web-based language sample analysis application for other languages, we researched the language technologies available for Lithuanian and Czech. Additionally, we researched language resources in Slovene, as this language shares similarities with Croatian but lacks language tests.

T3. Developing a survey for collecting information about language sample analysis in individual countries

We did not develop a survey to collect information about language sample analysis in individual countries. Instead, we used personal networks to approach individuals working in departments educating professionals who might use language sample analysis. We discovered that it is rarely used due to the lack of knowledge about the benefits and procedures, as well as the lack of appropriate (language-adapted and user-friendly) tools.

T4. Collecting information about language sample analysis in individual countries

We used personal networks to approach individuals working in departments educating professionals who might use language sample analysis. We discovered that it is rarely used due to the lack of knowledge about the benefits and procedures, as well as the lack of appropriate (language-adapted and user-friendly) tools.

T5. Developing strategies for the promotion of language sample analysis

We participated in several scientific conferences promoting language sample analysis. This was especially relevant for the case of Croatian, for which the web-based application for language-sample analysis was developed. Apart from presenting at conferences, we also had a public talk.

T6. Developing an open-source web-based application for language-sample analysis

An open-source web-based application for language-sample analysis was developed for Croatian. It can automatically recognize repetitions in transcripts, provide morphological tags for each word, parse the text into clauses, and determine the nature of the connective between the clauses. This enables fast and reliable calculation of lexical diversity measures (vocabulary diversity D, moving-average type-token ratio, lexical density), syntactic complexity measures (mean length of clause, mean length of communication unit, syntactic density). The application also incorporated a module for comparison (comparing individual results with a larger group of age-matched children).

Deliverables

(1) Publications

- Hržica, G., Liebeskind, C., Štrkalj Despot, K., Dontcheva-Navratilova, O., Kamandulytė-Merfeldienė, L., Košutar, S., Kramarić, M., Valūnaitė-Oleškevičienė, G. (2022). *Morphological Complexity of Children Narratives in Eight Languages*. Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022) / Calzolari, Nicoletta, Béchet, Frédéric, Blache, Philippe et al. (ur.). Pariz: European Language Resources Association (ELRA), 4729-4738.
- Košutar, S., Karl, D., Kramarić, M., Hržica, G. (2022). *Automatic Text Analysis in Language Assessment: Developing a MultiDis Web Application*. Proceedings of the Conference on Language Technologies and Digital Humanities / Fišer, Darja, Erjavec, Tomislav (Eds.). Ljubljana: Institute of Contemporary History, 2022. str. 93-99
- Košutar, S., Kramarić, M., Hržica, G. (2022). *The relationship between narrative microstructure and macrostructure: Differences between six- and eight-year-olds*. Psychology of Language and Communication, 26 (2022), 1; 126-153. doi: 10.2478/plc-2022-0007
- Košutar, S., Kramarić, M., Hržica, G. (2022). *Age-related differences in the expression of causal relationships during narrative production of Croatian children*. Rasprave Instituta za hrvatski jezik i jezikoslovlje, 48, 1; 327-347. doi: 10.31724/rihjj.48.1.15

(2) Conference/workshop presentations

- Hržica, Gordana, Liebeskind, Chaya, Štrkalj Despot, Kristina, Dontcheva-Navratilova, Olga, Kamandulytė-Merfeldienė, Laura, Košutar, Sara, Kramarić, Matea, Valūnaitė Oleškevičienė, Giedrė. Morphological Complexity of Children Narratives in Eight Languages. Language Resources and Evaluation Conference (LREC 2022), Marseille, France, 20.07.2022-25.07.2022
- Hržica, Gordana, Košutar, Sara, Karl, Dario, Kramarić, Matea. Selection, Implementation and Testing of Language Sample Analysis Measures for the Web-Based Application MultiDis. LLOD Approaches for Language Data Research and Management LLODREAM2022: International Scientific Interdisciplinary Conference / Autorių kolektyvas (ur.). Vilnius: Mykolo Romerio universitetas, 2022. str. 40-41.
- Košutar, Sara, Karl, Dario, Kramarić, Matea, Hržica, Gordana. Automatic Text Analysis in Language Assessment: Developing a MultiDis Web Application. Proceedings of the Conference on Language Technologies and Digital Humanities / Fišer, Darja, Erjavec, Tomislav (ur.). Ljubljana: Institute of Contemporary History, 2022. str. 93-99
- Hržica, Gordana, Bošnjak Botica, Tomislava, Košutar, Sara. More than just 'adding the -ed': Can we predict verb overgeneralizations in morphologically rich languages?. 20th International Morphology Meeting (Dedicated to the memory of Ferenc Kiefer). Budimpešta, Mađarska, 01.09.2022-04.09.2022

- Hržica, Gordana, Karl, Dario, Košutar, Sara, Kramarić, Matea. Automatska analiza jezičnih uzoraka uporabom aplikacije MultiDis 'Automatic analysis of language samples using application MultiDis'. Child and Languages Today, Croatia, Osijek, 16.09.2021-18.09.2021
- Hržica, Gordana, Karl, Dario, Košutar, Sara, Kramarić, Matea. Automatic analysis in language assessment. Language and Migrations (conference of the Croatian Applied Linguistics Society). Croatia, Osijek, 9.09.2021-11.09.2021
- Košutar, Sara, Kramarić, Matea, Hržica, Gordana. Age-related differences in the production of causal relations: evidence from narratives of Croatian children. International online conference Expressing causality in L1 and L2. Lublin, Poland, 20.05.2021-21.05.2021
- Košutar, Sara, Hržica, Gordana. Frequency and semantics of connective and in children's narrative discourse. Conference of the International Association for the Study of Child Language (IASCL Conference). USA, 15.07.2021-23.07.2021

(3) Datasets & other resources

- Compiled database of children language samples (multilingual)
- Compiled database language samples of children of different ages (Croatian)
- Additionally collected transcripts: Lithuanian and Croatian
- Language technologies adapted to analyses of child-spoken language
- Web-base application for language sample analysis in Croatian

Future work

Web-based Application for Analyzing Language Samples

- Exploring the Application:
  - Validation studies (in progress for syntactic complexity)
- Developing New Modules:
  - Integrating a module for speech-to-text (in progress)
  - Integrating a module for automatic error recognition
- Developing the Application for Other Languages:
  - Developing the version for Lithuanian
  - Exploring possibilities to develop applications for languages typologically similar to Croatian
- Expanding the Database needed for comparisons of individual results with group results

Research

- Development and validation of new measures of language sample analysis
- Adapting existing measures to new languages

2.1.3 UC4.1.3 Use Case in Acquiring RDF Relations with Neural Language Models

Coordinator: Dagmar Gromann (University of Vienna)

Overview

This use case targets a creative utilization of pre-trained neural language models for the acquisition of RDF relations across natural languages. This can be done in the form of question answer, e.g. When was X born? where the response represents the tail entity of the head entity X with predicate born in. Another option to acquire relations is to use the analogy task a is to be as c is to d, e.g. A buffalo is to bovid as bee is to __ where the models are asked to predict d, which in this case is insect. The idea is to provide a highly multilingual dataset in order to provide a benchmark for multilingual neural relation acquisition from neural language models, which includes low-resource languages, such as Albanian, Bambara, and Slovakian, and different scripts, such as Hebrew. Between 22 to 25 NexusLinguarum members from Working Groups 2, 3 and 4 participated in the regular telcos that took place approximately once a month.

Resources, methods, tools/technologies, languages used

As a major outcome of this use case, we generated a dataset of lexical semantic relations that is aligned across 15 languages, more are currently being processed and prepared. To be able to provide such a huge dataset, we decided to select an existing widely used dataset called Bigger Analogy Test Set (BATS)[1] and translate the existing proportion of lexical semantic relation data to as many languages as possible, benefiting from the highly multilingual and multicultural nature of NexusLinguarum. To this end, we had to provide translation guidelines so that the method for preparing the multilingual versions is identical. Since the dataset contains more than 3,000 pairs of words related by hypernymy, hyponymy, meronymy, antonymy, and synonymy, where possible more than one first-language speaker helped with the translation. To evaluate whether the translation guidelines indeed helped provide a similar translation result, each translator had to translate two sets from each relation type. On the basis of these parts translated by each translator, we calculated the agreement, which was different across languages. We found that factors such as the tendency to prefer loan words and age had a substantial effect on the deviation of translation decisions.

[1] https://vecto.space/projects/BATS/

The final dataset currently consists of 15 languages, which are Albanian, Bambara, Croatian, German, modern Greek, Hebrew, Italian, Latvian, Macedonian, Portuguese, Romanian, Spanish, Slovakian, and Slovenian. Some further languages, such as Ukrainian, are currently still being prepared. To show how to use this multilingual benchmark, we conducted experiments with the multilingual masked language model XLM-R and the generative open-source language model BLOOM on the typical analogy task introduced above as well as an innovative analogy-based translation task, where the objective is to predict the same tail entity in a different language, e.g. bee is to insect (EN) as Biene ist zu ____ (DE) where the correct answer would be the German translation of insect, which is Insekt. Both the dataset and the experiments have been presented in a publication that was recently accepted at the LREC-COLING 2024 conference and will be presented there in May 2024.

Roadmap, workflow and milestones

Given that this use case started quite late (March 2022), since the idea came up within the discussions in other NexusLinguarum contexts, most of the initial objectives and milestones were achieved. The only thing that could have benefited from more time are the experiments with neural language models, where we could to this date only publish a proportion of the designed and conducted experiments, i.e., only on the analogy-based tasks, whereas we also experimented with relation classification and intend to experiment with question-answering. In addition, we have started different RDF-based representation formats of the dataset, which is to be published before the end of the action in the format called Cross-Linguistic Data Formats (CLDF)[2].

Deliverables

(1) Publications

- Gromann, Dagmar, Hugo Gonçalo Oliveira, Lucia Pitarch, Elena-Simona Apostol, Jordi Bernad, Eliot Bytyçi, Chiara Cantone, Sara Carvalho, Francesca Frontini, Radovan Garabik, Jorge Gracia, Letizia Granata, Fahad Khan, Timotej Knez, Penny Labropoulou, Chaya Liebeskind, Maria Pia di Buono, Ana Ostroški Anić, Sigita Rackevičienė, Ricardo Rodrigues, Gilles Sérasset, Linas Selmistraitis, Mahammadou Sidibé, Purificação Silvano, Blerina Spahiu, Enriketa Sogutlu, Ranka Stanković, Ciprian-Octavian Truică, Giedrė Valūnaitė Oleškevičienė, Slavko Zitnik and Katerina Zdravkova (2024). *MultiLexBATS: Multilingual Dataset of Lexical Semantic Relations*. Accepted at LREC-COLING 2024.

[2] https://cldf.clld.org/v1.0/terms.rdf

- Garabík, Radovan; Ostroški Anic, Ana; Rackevičienė, Sigita; Valūnaitė Oleškevičienė, Giedrė; Selmistraitis, Linas and Utka, Andrius (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*. 12-15 September 2023, Vienna, Austria. NOVA CLUNL, Portugal. Ed. Sara Carvalho *et al*.: NOVA CLUNL, 2023. ISBN 9789895408153. p. 402-409. DOI: 10.34619/srmk-injj.

(2) Dataset

- Dataset – MultiLexBATS – of lexical semantic relations aligned across 15 languages (more are currently being processed and prepared)

Future work

Another publication on utilizing the dataset for further translation acquisition experiments is currently underway. Furthermore, the plan to continue this work is to use the already created CLDF representation to provide further insights into the similarities and differences across languages, e.g. lexical gaps, loan words, cognates, etc. Furthermore, we intend to intensify and extend the existing experiments on neural relation acquisition across languages and foresee further publications on this topic.

2.2 Task 4.2 Use Cases in Humanities and Social Sciences

Task Leaders

- Mietta Lennes, University of Helsinki (linguistics)
- Jouni Tuominen, University of Helsinki (computational)

Use Cases

- UC 4.2.1: Use Case in Humanities
- UC 4.2.2: Use Case in Social Sciences

Overview

The task 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 diachronic analysis in multilingual corpora that involve various time spans and text genres. It investigates the application of diachronic word embedding to a selection of datasets (in Latin, Hebrew, Lithuanian and French) and proposes a representation of the results as a linked data ontology. 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, the use cases advanced to the phases of conceptualization of the methodology for analysing change in word meaning over time and modelling it as linguistic linked open data (UC4.2.1) and method development for detection, extraction and semantic classification of discourse markers (UC4.2.2). The main outcomes of UC4.2.1 are the LLODIA model and a proof of concept that includes a sample of records built with data extracted from the selected datasets and dictionaries, as well as examples of dedicated queries and timelines. The main outcomes of UC4.2.2 include a discourse marker vocabulary of multiword expressions, a parallel corpus in several languages, and transformer ML models trained with the parallel corpus to predict the availability of discourse markers in unseen contexts.

The use case coordinators have organized activities, such as joint publications, workshops (Discourse studies and linguistic data science: Addressing challenges in interoperability, multilinguality and linguistic data processing – DiSLiDaS 2022 & 2023) and STSMs (UC4.2.2), as well as searched and applied for funding opportunities regarding future research projects. The use cases have identified collaboration possibilities with other WGs.

2.2.1 UC4.2.1 Use Case in Humanities

Coordinator: Florentina Armaselu (University of Luxemburg)

Overview

UC4.2.1 explores how linguistic data science can be used for diachronic analysis in multilingual corpora that involve various time spans and text genres. The use case investigates the application of diachronic word embedding to a selection of datasets (in Latin, Hebrew, Lithuanian, French and Romanian) and proposes a representation of the results as a linked data ontology. The LLODIA model, conceived for this purpose, uses linguistic linked open data (LLOD) and other Semantic Web formalisms, such as OntoLex(-FrAc), RDF, OWL and Dublin Core, to combine corpus- and dictionary-based evidence and encode cross-language, temporal and spatial relations derived from diachronic analysis. This type of resource is intended to provide contextual snapshots of word meaning across various temporal units, and linguistic and cultural spaces.

Observations from corpus processing, such as lists of most similar words to a form computed through embedding techniques and corpus citations from a certain time interval, can be linked to dictionary senses and their attestations, and to comparable representations in other languages that may involve translation or etymological relations. Specific information about the publication date and place encoded within the interconnected resources represents another way of tracing the evolution or circulation of word forms and their meaning across time and space. The main outcomes are the LLODIA model and a proof of concept that includes a sample of records built with data extracted from the selected datasets and dictionaries, as well as examples of dedicated queries and timelines. The use case involved an average number of 10 participants within regular monthly meetings (39 so far) and participation in joint workshops, conferences, and publications.

Resources, methods, tools/technologies, languages used

The core dataset of the use case includes texts in a variety of genres (literary, religious, technical, philosophical, historiographical, everyday life) and time periods:

- LatinISE (Latin, 2nd c. BC - 20th c. CE)
- Responsa (Hebrew, 11th - 21st c.)
- Sliekkas (Lithuanian, 16th - 18th c.)
- RoDICA (Romanian, half of the 19th - early 21st c.)
- BnL Open Data collection of monographs (French selection, 1690-1918).

The dictionary set used as reference and attestation source consists of mono- or multilingual resources, such as Portail lexical CNRTL for French, Charlton T. Lewis, Charles Short, A Latin Dictionary and latin-bert for Latin, Milog for Hebrew, LIETUVIUZODYNAS.lt and Lietuvių kalbos žodynas for Lithuanian, DEXonline for Romanian, and Wiktionary.

We used static diachronic word embeddings, computed via gensim word2vec (Mikolov et al., 2013; Rehurek and Sojka, 2010) for French and Hebrew, fastText (Bojanowski et al., 2017) for Latin and Lithuanian, and word2vec and ELMo (Truică et al., 2023) for Romanian.

For LLOD modelling, we utilised the RDF-XML format and tools such as Oxygen XML Editor and Protégé. Generative AI agents, such as ChatGPT-3.5, Chat-GPT-4 and Microsoft Copilot have been also tested for tasks related to exploration and assistance in generating LLOD representations and alignment of static word embedding results with dictionary entries.

Roadmap, workflow and milestones

(1) The recent technological advance in large language models and generative AI has led us to test the use of some of these technologies in the phase of LLOD pre-modelling and generation of our workflow.

(2) For the proof of concept of the LLODIA model to be published on the Web, we have identified and tested a set of semantic fields (e.g., culture, politics, economics, social and everyday life, family, health, religion, history) and target terms for each of the studied languages and analysed datasets. However, given time constraints, the usage examples included in the proof of concept contain a smaller set processed through manual and semi-automatic modelling.

Deliverables

(1) Publications

- Armaselu, Florentina. Chiarcos, Christian. McGillivray, Barbara. Khan, Anas Fahad. Truica, Ciprian-Octavian. Valūnaitė-Oleškevičienė, Giedrė. Liebeskind, Chaya. Apostol, Elena-Simona. Utka, Andrius. 2023. Towards a Conversational Web? A Benchmark for Analysing Semantic Change with Conversational Knowledge Bots and Linked Open Data. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 340–346, Vienna, Austria. NOVA CLUNL, Portugal.
- Armaselu, Florentina. McGillivray, Barbara. Liebeskind, Chaya. Valūnaitė-Oleškevičienė, Giedrė. Utka, Andrius. Gifu, Daniela. Khan, Anas Fahad. Apostol, Elena-Simona. Truica, Ciprian-Octavian. 2023. Workflow Reversal and Data Wrangling in Multilingual Diachronic Analysis and Linguistic Linked Open Data Modelling. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 410–416, Vienna, Austria. NOVA CLUNL, Portugal.

(2) Dataset

The LLODIA model and proof of concept is published on the Nexus Linguarum's GitHub repository: https://github.com/nexuslinguarum/LLODIA.

(3) Other deliverables

Use Case in Humanities. Linguistic Linked Open Data for Diachronic Analysis (LLODIA). Poster presented at the 6th Plenary Nexus Meeting, Athens, 20 March 2024. (Appendix 1)

Future work

We have two submissions accepted at the workshops DLnLD: Deep Learning and Linked Data (21 May 2024) and the 9th Workshop on Linked Data in Linguistics: Resources, Applications, Best Practices (25 May 2024) co-located with the LREC-COLING 2024:

- Armaselu, Florentina, Chaya Liebeskind, Paola Marongiu, Barbara McGillivray, Giedre Valunaite Oleskeviciene, Elena-Simona Apostol, Ciprian-Octavian Truică, and Daniela Gifu. "LLODIA: A Linguistic Linked Open Data Model for Diachronic Analysis." In *9th Workshop on Linked Data in Linguistics: Resources, Applications, Best Practices, Co-Located with the LREC-COLING 2024 Conference*. Turin, Italy, 2024.
- Armaselu, Florentina, Chaya Liebeskind, and Giedre Valunaite Oleskeviciene. "Self-Evaluation of Generative AI Prompts for Linguistic Linked Open Data Modelling in Diachronic Analysis." In *DLnLD: Deep Learning and Linked Data Workshop, Co-Located with the LREC-COLING 2024 Conference*. Turin, Italy, 2024.

We will resubmit our paper "Multilingual Word Embedding and Linguistic Linked Open Data for Tracing Semantic Change" to the *Rasprave* journal by 25 April 2024.

2.2.2 UC4.2.2 Use Case in Social Sciences

Coordinator: Mariana Damova (Mozaika Ltd, Bulgaria)

Overview

Speaker attitude detection is important for processing survey data as such data provide a valuable source of information and research for different scientific disciplines. Survey data offer evidence about particular language phenomena and public attitudes to present a broader picture of the clusters of social attitudes. In this regard, attitudinal discourse markers (DM) play a central role in the sense that they are pointers to the speaker's attitudes. The use case focused on the process of constituting a multilingual corpus, creating an annotation schema of discourse relations for marking the discourse markers, representing text containing DMs as Linked Data using OWL ontology, and applying machine learning transformer models to predict their appearance in unknown texts.

To meet these research goals, we created a parallel corpus containing data from 10 languages, namely English, Latvian, Hebrew, Bulgarian, European Portuguese, Polish, Romanian, Macedonian, German and Italian (prepared for CLARIN publication), using the publicly available TED Talk transcripts. The initial step was the manual annotation of 2,428 English-Bulgarian-Lithuanian aligned sentences containing the multiword expressions (MWE) as discourse markers or content expressions (0, 1). We trained the annotated corpora with transformer deep learning architectures (like XML-Roberta) and obtained models for prediction of the presence or absence of DMs in text with a good accuracy of 80-90%. We also applied a language-agnostic deep learning architecture (La-BSE) trained on the English annotated text onto the entire parallel corpus of 10 languages. Later in the project, we carried out an evaluation and validation of the performance of the language agnostic model on the 10 languages and found out that the performance was close to perfect.

Further, to follow the planned two-step annotation approach and account for the semantic and communicational role of DMs in text, we adapted the ISO-annotation schema, annotated a chunk of the parallel corpus of 10 languages, while also developing a Linked data representation of the texts annotated with this annotation schema and converted into an OWL ontology. We produced 9 publications (conference and workshop papers), and have a pending submission of a journal paper, covering the linguistic side of our research.

Resources, methods, tools/technologies, languages used

- Publicly available TED talk transcripts
- Parallel corpus in 10 languages with English as pivot language, resulting in 9 bilingual parallel corpora – English-Latvian, English-Hebrew, English-Bulgarian, English-European Portuguese, English-Polish, English-Romanian, English-Macedonian, English-German, English-Italian
- Vocabulary of discourse markers (DM) in 10 languages – English, Latvian, Hebrew, Bulgarian, European Portuguese, Polish, Romanian, Macedonian, German, Italian

- Manually annotated corpus of English-Latvian-Bulgarian for the presence or absence of discourse markers
- Trained and validated language models (XML-Roberta) for predicting the presence or absence of discourse markers in unseen text in English, Latvian and Bulgarian
- Trained language agnostic models (LA-BSE) on English for the presence or absence of discourse markers
- Corpora in 9 languages – Latvian, Hebrew, Bulgarian, European Portuguese, Polish, Romanian, Macedonian, German, Italian – produced with the language agnostic models, annotated with the presence or absence of discourse markers
- Validation of the performance of the language agnostic models
- ISO-based annotation schema for discourse markers in text
- OWL ontology based on the annotation schema
- Linguistic linked data representation of text, based on the developed OWL ontology
- Parallel corpus in 10 languages with ISO-based annotations

Roadmap, workflow and milestones

Table 2. UC4.2.2 planned roadmap, workflow and milestones

Deliverables

(1) Publications

- Christian Chiarcos, Purificação Silvano, Mariana Damova, Giedre Valunaite Oleškeviciene, Chaya Liebeskind, Dimitar Trajanov, Ciprian-Octavian Truica, Elena-Simona Apostol, Anna Baczkowska. *Building an Owl-Ontology for Representing, Linking and Querying SemAF Discourse Annotations* In: Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje, Vol. 49 No. 1, 2023.
- Mariana Damova, Kostadin Mishev, Giedre Valunaite Oleškeviciene, Chaya Liebeskind, Purificação Silvano, Dimitar Trajanov, Ciprian-Octavian Truica, Elena-Simona Apostol, Christian Chiarcos, Anna Baczkowska**.** *Validation of Language Agnostic Models for Discourse Marker Detection*. In Proceedings of LDK2023, Vienna, Austria, September 2023.
- Emma Angela Montechiari, Kostadin Mishev, Stanko Stankov and Mariana Damova. *Machine Learning Methods for Discourse Marker Detection in Italian*. In Proceedings of Workshop on Deep Learning and Neural Approaches for Linguistic Linked Data (DL4LD), NexusLinguarum Workshop. Vilnius, Lithuania, September 2022.

- Christian Chiarcos, Purificação Silvano, Mariana Damova, Giedre Valunaite Oleškeviciene, Chaya Liebeskind, Dimitar Trajanov, Ciprian-Octavian Truica, Elena-Simona Apostol, Anna Baczkowska. *An OWL Ontology for ISO-based Discourse Marker Annotation*. In Proceedings of "LLOD approaches for language data research and management" (LLODREAM2022), NexusLinguarum Conference. Vilnius, Lithuania, September 2022.
- Barbara Lewandowska-Tomaszczyk, Mariana Damova. *(Common) ground and Discourse Development Prediction Associated with the Role of Intonation in the Interpretation of Communicative Connectives*. SLE 2022, Bucharest, Romania, August 2022.
- Purificação Silvano, Mariana Damova, Giedre Valunaite Oleškevicien, Chaya Liebeskind, Chaya Liebeskind, Christian Chiargos, Dimitar Trajanov, Ciprian-Octavian Truica, Elena-Simona Apostol, Anna Baczkowska. *ISO-based Annotated Multilingual Parallel Corpus for Discourse Markers*. Poster. LREC 2022, Marseille, France, June 2022.
- Purificação Silvano, Mariana Damova. *ISO-DR-core plugs into ISO-dialogue acts for a crosslinguistic taxonomy of discourse markers*. DiSLiDaS 2022 workshop, NexusLinguarum, Jerusalem, Israel, May 2022.
- Kostadin Mishev, Mariana Damova, Giedre Valunaite Oleškevicien, Chaya Liebeskind, Chaya Liebeskind, Dimitar Trajanov, Purificação Silvano, Ciprian-Octavian Truica, Elena-Simona Apostol, Christian Chiargos. *Evaluation of Cross-Lingual Methods for Discourse Markers Detection*. DiSLiDaS 2022 workshop, NexusLinguarum, Jerusalem, Israel, May 2022.
- Giedre Valunaite Oleskevicienė, Chaya Liebeskind, Dimitar Trajanov, Purificação Silvano, Christian Chiarcos, Mariana Damova. *Speaker Attitudes Detection through Discourse Markers Analysis*. In: Proceedings of Workshop "Deep Learning and Neural Approaches for Linguistic Data", NexusLinguarum, Skopje, October 2021.

(2) Datasets & other resources

- Parallel corpus in 10 languages with English as pivot language, resulting in 9 bilingual parallel corpora – English-Latvian, English-Hebrew, English-Bulgarian, English-European Portuguese, English-Polish, English-Romanian, English-Macedonian, English-German, English-Italian
- ISO-based annotation schema for DMs in text
- OWL ontology based on the annotation schema
- Trained and validated language models (XML-Roberta) for predicting the presence or absence of discourse markers in unseen text in English, Latvian and Bulgarian
- Dataset of 10 languages with predicted with language agnostic models discourse marker presence and absence
- Validated performance of language agnostic models trained on English for 9 unseen languages

(3) Other deliverables

Use Case in Social Sciences. Poster presented at the 6th Plenary Nexus Meeting, Athens, 20 March 2024. (Appendix 2)

Future work

For future research, we foresee social attitudes detection, opinions, and speaker's attitude as the next steps.

2.3 Task 4.3 Use Cases in Technology

Task Leaders

- Tilia Ellendorff, University of Zurich (linguistics)
- Dimitar Trajanov, University ss. Cyril and Methodius – Skopje (computational)

Use Cases

- UC 4.3.1: Use Case in Cybersecurity
- UC 4.3.2: Use Case in FinTech

Overview

This task leveraged recent advancements in NLP, automatic term extraction, text analytics, and sentiment analysis. The objective was to integrate and test existing open-source components across various Information and Communication Technology (ICT) and business contexts. The tasks involved in this project encompassed conducting state-of-the-art analyses, eliciting requirements and defining use cases, compiling corpora, extracting terms and linking them semantically, classifying documents, and evaluating Natural Language Processing (NLP) tools in diverse scenarios.

Two specific Use Cases were prioritized: Cybersecurity and Financial Technology (FinTech). The Cybersecurity Use Case aimed to compile a bilingual cybersecurity termbase that included terminological data extracted from corpora using deep learning systems. The FinTech Use Case focused on developing domain-specific sentiment analysis models designed to extract actionable insights from financial news efficiently.

2.3.1 UC4.3.1 Use Case in Cybersecurity

Coordinator: Sigita Rackevičienė (Mykolas Romeris University, Lithuania)

Overview

The use case aimed to develop a methodology for the compilation of a bilingual (English-Lithuanian) termbase for the domain of cybersecurity (CS), using deep learning systems and LLOD principles. The pursuit was motivated by two factors: the under-resourced state of the Lithuanian language, and the ever-growing relevance of cybersecurity and its terminology in contemporary discourse.

The work started with the compilation of a bilingual parallel corpus and a bilingual comparable corpus, which enabled the extraction of terms from English texts and their Lithuanian translations, as well as from original English and Lithuanian texts that covered similar topics and were written in similar discourses. The subsequent stage involved manual annotation of training corpora (gold standard) and the training of deep learning systems, which were then employed for automatic extraction of cybersecurity terminology and for aligning English and Lithuanian terminological counterparts. The resulting terminological datasets underwent manual revision, and a selection of terms, deemed pertinent for the termbase, was concluded with the collaboration of a cybersecurity expert. The selected terms were grouped into concept sets, with each set representing a distinct concept, and further organised into thematic categories and subcategories on the basis of a conceptual model of the cybersecurity domain developed for the purposes of the use case.

The terminological dataset was enriched with additional textual and numerical data essential for a comprehensive termbase, including concept definitions, contextual term usage examples, and term frequencies in the corpora. The publicly open termbase, titled *Lithuanian-English Cybersecurity Termbase* / *Lietuvių-anglų kalbų kibernetinio saugumo terminų bazė*, was developed on Terminologue platform administered by Dublin City University. In the ongoing final stage, the bilingual corpora datasets (in txt and tmx formats) and the termbase dataset (in tbx format) are currently being converted into a linkable format to link them to each other and integrate them into the Linked Language Open Data (LLOD) Cloud.

Number of participants: six (Sigita Rackevičienė, Andrius Utka, Liudmila Mockienė, Aivaras Rokas, Max Ionov, Christian Chiarcos).

Approximate number of meetings: one online meeting a week.

Resources, methods, tools/technologies, languages used

Languages: English, Lithuanian.

Resources for the corpus datasets used for terminology extraction:

- international and national legislative documents on cybersecurity issues,
- public documents of international and national cybersecurity institutions,
- academic literature in the field of cybersecurity,
- specialised and mass media articles covering cybersecurity topics.

Resources for additional data:

- cybersecurity glossaries,
- ISO standards on information security.

Methods and tools:

- parallel and comparable corpora building methodology (tools used: various search engines for search of texts, bilingual text aligners for alignment of parallel texts, English and Lithuanian morphological analysers for POS-tagging of texts, corpus building and query toolkits);
- small-scale training corpora (gold standard) development methodology (tools used: bilingual terminology annotation software developed for the purposes of the use case);
- automatic bilingual terminology extraction methodology (tools used: deep learning systems developed for the purposes of the use case);
- termbase compilation methodology (tools used: cloud-based instance of terminology management software);
- linguistic data linking methodology (tools used: converters of tbx to rdf and tmx to rdf developed for the purposes of the use case).

Roadmap, workflow and milestones

Figure 2. Roadmap with milestones for the development of a bilingual cybersecurity (CS) termbase

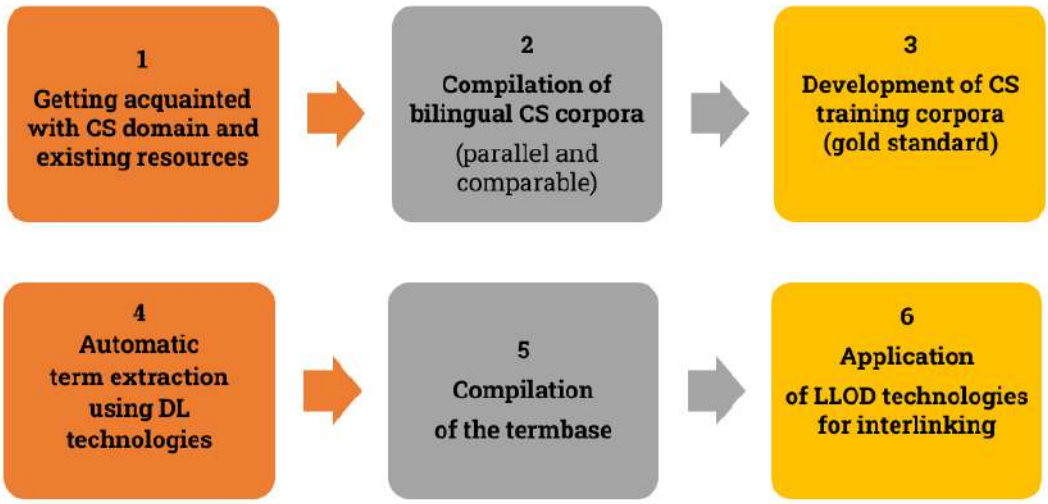
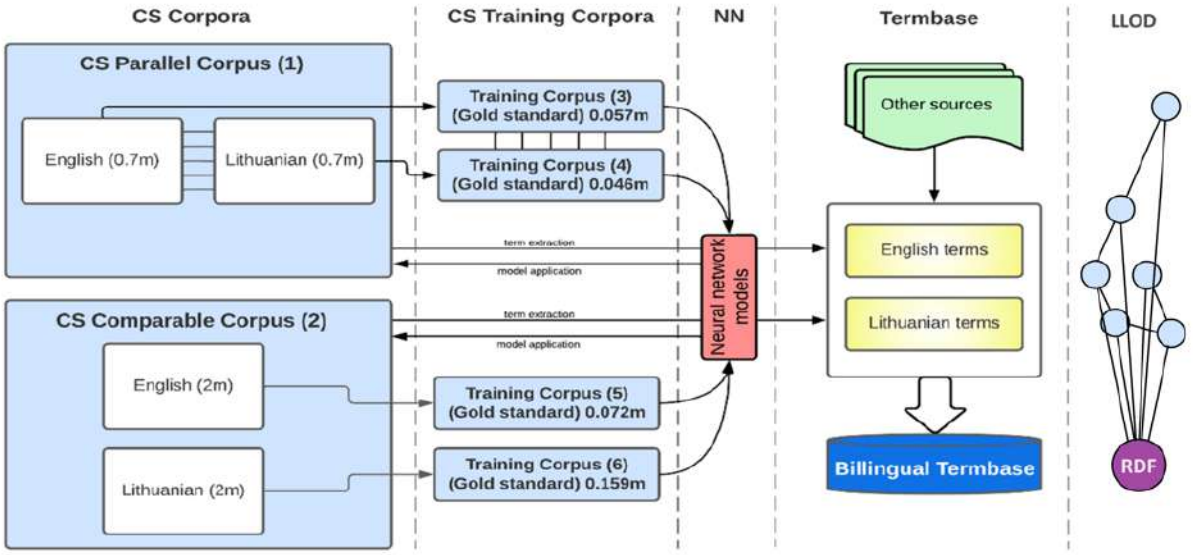


Figure 3. Workflow of the collection of cybersecurity (CS) data for the termbase



Deliverables

(1) Publications:

- Rackevičienė, Sigita; Utka, Andrius; Bielinskienė, Agnė; Mockienė, Liudmila. *Lithuanian-English cybersecurity termbase: principles of data collection and structuring* // Rasprave. 2023, vol. 49, iss. 2, p. 1-24. DOI: 10.31724/rihjj.49.2.12.
- Rackevičienė, Sigita; Utka, Andrius. *Developing Training Corpora for Automatic Extraction of Cybersecurity Terminology* // TOTh 2022. Terminologie & Ontologie: Théories et Applications: Actes de la conférence, 2 & 3 juin 2022. Université Savoie Mont Blanc, 2023, p. 75-95.
- Rackevičienė, Sigita; Utka, Andrius; Bielinskienė, Agnė; Rokas, Aivaras. *Distribution of Terms Across Genres in the Annotated Lithuanian Cybersecurity Corpus* // Respectus philologicus, 2022, vol. 41(46), p. 26-42. DOI: 10.15388/RESPECTUS.2022.41.46.105.
- Utka, Andrius; Mockienė, Liudmila; Laurinaitis, Marius; Rackevičienė, Sigita; Rokas, Aivaras; Bielinskienė, Agnė. *Building of Parallel and Comparable Cybersecurity Corpora for Bilingual Terminology Extraction* // Selected Papers from the CLARIN Annual Conference 2021 Virtual Event, 2021, 27–29 September. Linköping University Electronic Press, 2022, p. 126-138. DOI: 10.3384/ecp18912
- Rackevičienė, Sigita; Mockienė, Liudmila; Utka, Andrius; Rokas, Aivaras. *Methodological Framework for the Development of an English-Lithuanian Cybersecurity Termbase* // Studies about Languages. 2021, no. 39, p. 85-92. DOI: 10.5755/j01.sal.1.39.29156.
- Rokas, Aivaras; Rackevičienė, Sigita; Utka, Andrius. *Automatic Extraction of Lithuanian Cybersecurity Terms Using Deep Learning Approaches* // Human Language Technologies – The Baltic Perspective. Proceedings of the Ninth International Conference Baltic HLT 2020. IOS Press, 2020, p. 39-46. DOI: 10.3233/FAIA200600.

(2) Datasets:

- Utka, Andrius; Rackevičienė, Sigita; Rokas, Aivaras; Bielinskienė, Agnė; Mockienė, Liudmila and Laurinaitis, Marius, 2022, **English-Lithuanian Parallel Cybersecurity Corpus** - DVITAS, CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/46 .
- Utka, Andrius; Rackevičienė, Sigita; Rokas, Aivaras; Bielinskienė, Agnė; Mockienė, Liudmila and Laurinaitis, Marius, 2022, **English-Lithuanian Comparable Cybersecurity Corpus** - DVITAS, CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/47 .
- **Lithuanian-English Cybersecurity Termbase** on Terminologue platform: https://www.terminologue.org/csterms/
- **Lithuanian-English Cybersecurity Termbase** dataset in tbx format: Utka, Andrius; Rackevičienė, Sigita; Bielinskienė, Agnė; Laurinaitis, Marius; Mockienė, Liudmila and Rokas, Aivaras, 2023, **Lithuanian-English Cybersecurity Termbase** v.0.1, CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/55 .

(3) Other deliverables

Use Case in Cybersecurity. Poster presented at the 6th Plenary Nexus Meeting, Athens, 20 March 2024. (Appendix 3)

Dissemination

The results of the use case have been presented in:

(1) WG4 meetings and NexusLinguarum general meetings

(2) LexicalaReview 29, July 2021

https://lexicala.com/wp-content/uploads/review/2021/an-overview-of-nexuslinguarum-use-cases-current-status-and-challenges.pdf

(3) Conferences and seminars:

- Presentation *Automatic Extraction of Lithuanian Cybersecurity Terms Using Deep Learning Approaches* delivered at the conference *Baltic HLT 2020: Human Language Technologies - the Baltic Perspective* (Kaunas, Lithuania, 2020) https://sitti.vdu.lt//hlt/
- Presentation *Terminology of Cyber Domain: Some Insights into the Conceptual and Linguistic Dimensions* delivered at the conference *Terminology – Heritage and Modernity* (Tbilisi, Georgia, 2020) https://ice.ge/of/?page_id=4780
- Presentation *Cybersecurity Terms in Lithuanian: How Do We Create Them and Which Formations Do We Prefer?* delivered at the seminar *Outside the Frames: New Challenges for Terminology Work* (Zagreb, Croatia, 2021) http://ihjj.hr/dika/wp-content/uploads/2021/03/outside-the-frames_final-program.pdf

- Presentation *Corpora for Bilingual Terminology Extraction in Cybersecurity Domain* delivered at the conference *CLARIN Annual Conference* (Netherlands, Utrecht, 2021) https://www.clarin.eu/event/2021/clarin-annual-conference-2021-virtual-event
- Presentation *Annotation of Cybersecurity Terminology: Methodology, Problems and Results* delivered at the conference *Scientific, Administrative and Educational Dimensions of Terminology* (Vilnius, Lithuania, 2021) http://lki.lt/en/4-oji-tarptautine-moksline-terminologijos-konferencija/
- Presentation *Developing Training Corpora for Automatic Extraction of Cybersecurity Terminology* delivered at the conference *TOTh 2022: Terminology & Ontology: Theories and Applications* (Chambery, France, 2022) http://toth.condillac.org/toth-2022
- Presentation *Terminology Studies at Mykolas Romeris University* delivered at the conference *Terminology and Specialized Knowledge Representation: New Perspectives on User Needs* (Rijeka, Croatia, 2022) https://www.jezik.hr/tesk2022/
- Presentation *Developing a Cybersecurity Termbase* delivered at the conference *LLOD Approaches to Language Data Research and Management, LLODREAM* (Vilnius, Lithuania, 2022) https://llodapproaches2022.mruni.eu/
- Presentation *Terminology of the Cybersecurity Domain: Formation Patterns of the Most Frequent English and Lithuanian Terms in the Bilingual ad hoc Corpora* delivered at the conference *Terminology – Heritage and Modernity* (Tbilisi, Georgia, 2022) https://ice.ge/of/?page_id=5900%20;%20%20https://bit.ly/3XBRVTP
- Presentation *Enhancing Interoperability for Under-Resourced Languages: A Case Study on Linking Lithuanian-English Data in the Cybersecurity Domain* delivered at UNIDIVE (COST Action CA21167) Workshop (Naples, Italy, 2024) https://unidive.lisn.upsaclay.fr/doku.php?id=meetings:general_meetings:2nd_unidive_general_meeting:abstracts#session_b

Future work

The use case team plans to extend cooperation with international colleagues in the W3C Ontology-Lexicon community and other relevant communities/networks. The cooperation will focus on modelling corpus and terminological datasets and enhancing their interoperability by application of LLOD technologies.

2.3.2 UC 4.3.2 Use Case in Fintech

Coordinator: Dimitar Trajanov (University ss. Cyril and Methodius – Skopje)

Overview

The Fintech use case encompasses two distinct subtasks, each aimed at refining sentiment analysis within the finance sector by employing natural language processing (NLP) technologies.

The first subtask, "Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers," embarks on a chronological investigation of over 100 models for sentiment analysis, tracing the evolution from lexicon-based methods through word and sentence encoders to the contemporary NLP transformers. This study underscores the superior performance of NLP transformers in capturing the semantics of financial texts, with distilled versions like Distilled-BERT and Distilled-RoBERTa offering a blend of high efficiency and reduced resource consumption suitable for production environments. Despite the constraints of a small dataset of only 2K sentences, the research reveals that these models can deliver expert-level sentiment analysis.

The second subtask, "Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (Xlex)," shifts focus towards developing a new methodology that merges the strengths of lexicon-based approaches with the advanced capabilities of transformer models. This approach, termed eXplainable Lexicons (Xlex), leverages transformers and Shapley Additive exPlanations (SHAP) to autonomously generate financial lexicons that extend beyond the benchmark Loughran-McDonald lexicon, reducing human effort in lexicon annotation and upkeep. The resultant Xlex framework surpasses traditional lexicons in sentiment analysis accuracy and offers enhanced interpretability and efficiency, crucial for real-time applications and systems with constrained computational resources.

Collaboration in this UC spanned across a diverse group of experts in natural language processing (NLP), finance, and computational linguistics, adopting a multidisciplinary approach that significantly enhanced our research and development processes.

The Fintech use case illustrates the dynamic landscape of sentiment analysis in finance, showcasing the integration of traditional techniques with modern NLP technologies to improve efficiency, accuracy, and applicability across various domains.

Resources, methods, tools/technologies, languages used

In the use case, the following methods and models were used: Lexicon-Based Approach, Word and Sentence Encoders, NLP Transformers, and explainable ML methods like SHAP.

The following Datasets were used for model development:

- Nasdaq (Version 2), https://www.kaggle.com/datasets/sidarcidiacono/news-sentiment-analysis-for-stock-data-by-company
- Financial phrase bank (Version 5), https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news?select=all-data.csv
- Sentfin (Version 3), https://www.kaggle.com/datasets/ankurzing/aspect-based-sentiment-analysis-for-financial-news
- Sem Eval (train and trial), https://alt.qcri.org/semeval2017/task5/index.php?id=data-and-tools
- https://bitbucket.org/ssix-project/semeval-2017-task-5-subtask-2/src/master
- FIQA (train), https://sites.google.com/view/fiqa/
- Financial Phrase Bank + FIQA (Version 4), https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis
- Loughran McDonald Dictionary, https://sraf.nd.edu/loughranmcdonald-master-dictionary/

Roadmap, workflow and milestones

The use case began with an in-depth review of sentiment analysis techniques, focusing on those applicable to finance. Following this, we started collecting and preprocessing a dataset comprising approximately 2,000 financial phrases. This preparation was essential to ensure the dataset's compatibility with various sentiment analysis models. The core of our implementation and analysis phase involved several key steps:

- Evaluation of lexicons, word, and sentence encoders for sentiment analysis.
- Evaluating NLP transformers like BERT and RoBERTa for advanced NLP analysis.
- Implementing lexicon-based models as a baseline.
- Developing the Xlex methodology to enhance financial lexicons automatically using transformers and SHAP for explainability.

Deliverables

(1) Publications

- Mishev, Kostadin, Ana Gjorgjevikj, Irena Vodenska, Lubomir T. Chitkushev, and Dimitar Trajanov. "Evaluation of sentiment analysis in finance: from lexicons to transformers." IEEE Access 8 (2020).
- Rizinski, Maryan, Hristijan Peshov, Kostadin Mishev, Milos Jovanovik, and Dimitar Trajanov. "Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (Xlex)." IEEE Access (2024).

(2) Other resources

- A comprehensive analysis of more than 100 models for sentiment analysis was conducted. The results and code were published on Git Hub.
- A new Xlex model for automating the process of dictionary creation was developed.

The source code of the models and analysis is open source and available on GitHub with an MIT license.

- Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers https://github.com/f-data/finSENT
- Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (Xlex). https://github.com/hristijanpeshov/SHAP-Explainable-Lexicon-Model

(3) Other deliverables

Use Case in FinTech. Poster presented at the 6th Plenary Nexus Meeting, Athens, 20 March 2024. (Appendix 4)

Future work

For future work, we aim to further the reach and application of our research by developing a comprehensive library based on the Xlex methodology. This work will involve designing a user-friendly interface and robust API that researchers and practitioners in the field of finance, as well as other domains, can easily integrate into their projects. The library will encapsulate the innovative aspects of Xlex, offering functionalities for automatic lexicon enhancement, sentiment analysis, and explainability features that leverage transformers and SHAP for detailed insights into model decisions.

2.4 Task 4.4 Use Cases in Life Sciences

Task leaders

- Ana Ostroški Anić, Institute of Croatian Language and Linguistics (linguistics)
- Marko Robnik-Šikonja, University of Ljubljana (computational)

Use Cases

- UC4.4.1: Use Case in Public Health
- UC4.4.2: Use Case in Pharmacology

Overview

The area of Life Sciences is too broad and heterogeneous to be covered as a whole. For that reason, T4.4 was constrained to a general overview and focused investigation of two important subtopics: Public Health and Pharmacology. Our investigation in the public health use case was targeted at the quality of life and particularly the COVID-19 pandemic, which shaped life during the Action's duration. UC4.4.1 covered public health topics predominantly within news media, social media, and parliamentary speech, using multilingual and cross-lingual approaches, including linguistic analyses and large language models. In the pharmacology use case, we investigated the application of natural language processing techniques in this domain. Here, we also used other sources of information, predominantly scientific literature on life sciences and its relationship with linked data.

The results are varied and include the forging of several previously non-existent research collaborations. Measurable results include high-quality joint publications, presentations in the public domain and scientific venues, as well as joint project proposals of Action participants.

2.4.1 UC4.4.1 Use Case in Public Health

Coordinators: Petya Osenova (Institute of Information and Communication Technologies, Bulgarian Academy of Science) and Marko Robnik-Šikonja (Faculty of Computer and Information Science, University of Ljubljana)

Overview

This use case targeted predominantly disease prevention and quality of life.

The main task was to cover tendencies within news media, social media and other media available for research in a cross-lingual setting. The main source of information was the COVID-19 pandemic situation. Another source of information was the scientific literature on life sciences as well as its relationship with linked data.

Resources, methods, tools/technologies, languages used

The resources used in our work were: available ontologies, corpora and lexical databases (such as terminological dictionaries). The approaches include machine learning, information extraction, and NLP processing pipelines but more specifically linked open data, embeddings, and knowledge graphs.

In the cross-lingual settings, we focused on non-English languages, such as Slovenian, Bulgarian, Croatian, Portuguese, Romanian, Czech, Spanish, Lithuanian, Spanish, French, etc.

The number of participants was around 10, which means that our potential was limited. However, the results show that despite the small number of persons involved, important findings were reached.

Roadmap, workflow and milestones

We started with an informative survey of the SOTA in the selected topics, covering specific resources, methods, technologies and approaches. Based on that, we decided to focus on two topics:

- Researching varieties of metaphors related to the pandemic situation through the usage of ParlaMint 2.0 corpora as uploaded into the NoSketch concordancer, as well as the iInitial modelling of the metaphor frames into an ontology and its connection to the appropriate related lexica. The basis for the modelling was the paper of Despot, Kristina; Ostroški Anić, Ana. 2021. A War on War Metaphor: Metaphorical Framings in Croatian Discourse on Covid-19. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje* 47/1. 173–208. https://doi.org/10.31724/rihjj.47.1.6.
- Researching the parliamentary speech in several countries before and during the COVID-19 pandemic. Here ParlaMint 2.0 corpora were used as data sources. We detected and compared the most prominent topics appearing in the discourse for six parliaments, as well as the sentiment, emotions, and differences in language based on speakers' gender, age, and political orientation.

The ParlaMint corpora have been described in: Erjavec, Tomaž; Ogrodniczuk, Maciej; Osenova, Petya et al. 2022. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-021-09574-0.

We did not succeed in modelling all the metaphor frames for the languages envisaged. However, an initial ontology and its connection to the related lexica was created. Also, the presentation of Liudmila Mockienė on Lithuanian, the publications of Petya Osenova on Covid metaphors in the Bulgarian parliamentary corpus as well as the paper of (Despot, Kristina and Ostroški Anić, Ana 2021) and their observations on Croatian data ensure the basis for future research.

Deliverables

Initially, three deliverables had been planned, but given the change in the focus of activities, they have not been written in their planned form. The identification and description of related resources and tools were rather incorporated into the presentations.

Deliverables not planned, but achieved, are listed below.

(1) Publications

1. Osenova, Petya. 2021. Metaphors of Pandemia in the Bulgarian Parliamentary Data: a Corpus Survey. *Linguistic issues*, year II, issue 2. Todorova, Bilyana; Padareva-Ilieva, Gergana (eds.). 279–286. Blagoevgrad: University Publishing House "Neofit Rilski". // Осенова, Петя. 2021. Метафорите на пандемията в българската парламентарна реч: корпусно изследване. *Лингвистични проблеми* II/2. Тодорова, Биляна; Падарева-Илиева, Гергана (отг. Редактори). 279–286. Благоевград: Университетско издателство „Неофит Рилски"]
2. Kristian Miok, Encarnacion Hidalgo-Tenorio, Petya Osenova, Miguel-Angel Benitez-Castro, Marko Robnik-Sikonja. 2024. Multi-aspect Multilingual and Cross-lingual Parliamentary Speech Analysis. *Intelligent Data Analysis*. Pre-press at: https://content.iospress.com/articles/intelligent-data-analysis/ida227347
   Another link: https://arxiv.org/abs/2207.01054.
3. Osenova 2022: The pandemic of COVID 19 in the valency of its predicated: observations on a contemporary corpus of parliamentary speech. In: *BULGARIAN LANGUAGE, SUPPLEMENT*, 69 (2022), 113–121, Print ISSN: 0005-4283, Online ISSN: 2603-3372, doi: 10.47810/BL.69.22.PR.06 // Осенова 2022: Петя Осенова. ПАНДЕМИЯТА ОТ COVID-19 ВЪВ ВАЛЕНТНОСТТА НА СВОИТЕ ПРЕДИКАТИ: НАБЛЮДЕНИЯ ВЪРХУ СЪВРЕМЕНЕН КОРПУС ОТ ПАРЛАМЕНТАРНА РЕЧ. В: сп. *БЪЛГАРСКИ ЕЗИК, ПРИЛОЖЕНИЕ* 69 (2022), 113–121, Print ISSN: 0005-4283, Online ISSN: 2603-3372, doi: 10.47810/BL.69.22.PR.06
4. Markovikj, Marko; Dobreva, Jovana; Lucas, Mary; Vodenska, Irena; Chitkushev, Lou; Trajanov, Dimitar. 2022. Terminology and topics analysis of tweets related to the COVID-19 pandemic. Extended abstract in *Terminology and Specialized Knowledge Representation: New Perspectives on User Needs. Book of Abstracts*. Ostroški Anić, Ana; Grčić Simeunović, Larisa; Rajh, Ivanka (eds.). Institute of Croatian Language and Linguistics. Zagreb.

5. Declerck, Thierry. 2022. Towards a Multilingual ATC Ontology. Extended abstract in *Terminology and Specialized Knowledge Representation: New Perspectives on User Needs. Book of Abstracts*. Ostroški Anić, Ana; Grčić Simeunović, Larisa; Rajh, Ivanka (eds.). Institute of Croatian Language and Linguistics. Zagreb.

(2) Conference/workshop presentations

1. Kristina Despot, Liudmila Mockienė, Petya Osenova and Ana Ostroški Anić: Covid-19 and health related metaphors in Bulgarian, Croatian and Lithuanian ParlaMint corpora. Presentation at: *Terminology and Specialized Knowledge Representation: New Perspectives on User Needs* (TESK 2022), 9–10 June 2022, University of Rijeka, Faculty of Law.
2. Markovikj, Marko; Dobreva, Jovana; Lucas, Mary; Vodenska, Irena; Chitkushev, Lou; Trajanov, Dimitar. 2022. Terminology and topics analysis of tweets related to the COVID-19 pandemic. Presentation at: *Terminology and Specialized Knowledge Representation: New Perspectives on User Needs* (TESK 2022), 9–10 June 2022, University of Rijeka, Faculty of Law.
3. Marko Robnik-Šikonja, 2023. Challenges in explaining machine learning models for text. Invited lecture at SemDial – The 27th Workshop on the Semantics and Pragmatics of Dialogue, Maribor, Slovenia, 17 August 2023
4. Marko Robnik-Šikonja, 2023. Large language models for cross-lingual transfer. Invited lecture in 34th European Summer School in Logic, Language and Information, Ljubljana, Slovenia, 01 August 2023
5. Marko Robnik-Šikonja, 2020. Cross-Lingual Embeddings: A Babel Fish for Machine Learning Models. Keynote address at Artificial Intelligence and Natural Language Conference, AINL 2020, online, 08 October 2020

(3) Datasets

- Ontology of the COVID-related metaphorical frames

(4) Research project

Ana Ostroški Anić and Marko Robnik-Šikonja's project proposal *Automatic identification of semantic relations in figurative context in Croatian and Slovene*, which had been initiated in Task 4.4, was granted funding within the Slovene-Croatian research framework. The project started on 1 April 2023 and will finish on 31 March 2025.

(5) Other deliverables

Use Case in Public Health. Poster presented at the 6th Plenary Nexus Meeting, Athens, 20 March 2024. (Appendix 5)

Development of the ontology of the COVID-related metaphorical frames

Within the scope of this UC and supported by an STSM carried out by Liudmila Mockienė at the Department of AI and Language Technologies, IICT-BAS, in August 2021. Parliamentary Data from the ParlaMint project were processed in the NoSketch concordancer for Bulgarian, Croatian and Slovenian, thus aiming to analyse varieties of metaphors related to the current pandemic situation.

The aim was the extension of multilingual resources, i.e. carrying out research on COVID-19 and health-related metaphors based on the multilingual corpus ParlaMint, and the creation of a corpus covering the COVID-19 pandemic situation as presented in the Lithuanian news and social media. In addition, a corpus analysis of COVID-19 and health-related metaphors was envisaged.

The first step while working on ParlaMint-LT 2.0 was to create a subcorpus for analysing COVID-related metaphors. It was analysed further to establish the keyness (relevance) of the keywords in comparison with the rest of the ParlaMint-LT 2.0 corpus. Each keyword related to COVID (pandemija (pandemic), epidemija (epidemic), COVID, COVID-19, virusas (virus), koronavirusas (coronavirus), korona (corona), karantinas (quarantine)) was further analysed individually to extract the concordance lines which include specific context the keyword is used in. 1,305 concordance lines were extracted for the qualitative analysis. Each concordance line was analysed to establish whether the keyword was used metaphorically or not. In total, 301 cases were marked as metaphorical, which is 23% of all cases. All metaphoric phrases were 'lemmatised' and translated into English.

Next, the frame of the metaphor was established. The frames found in the data are PERSONIFICATION (22%), EVENT STRUCTURE > CAUSES ARE FORCES (20%), COMBAT > WAR (17%), REIFICATION (13%), CONTROL (10%), DISASTER > FIRE (4%), DISASTER > CRISIS (3.3%), DISASTER > THREAT (2.4%), DISASTER > WATER (2.3%), EVENT STRUCTURE > ACTION IS MOTION (2%), DISASTER (1.6%), DANGER (0.6%), PARTNERSHIP (0.3%).

Sample entry from the ontology[3]

```
%%%%%%% Lexical Entries
nexusi:LexicalEntry1 a nexuso:LexicalEntry ;
   nexuso:lexicalForm [nexuso:writtenRep "pandemijos įveikimas"@lt ; nexuso:writtenRep
"overcoming/ beating the pandemic"@en] ;
   nexuso:frame nexuso:Combat ;
   nexuso:frame nexuso:War .
```

[3] The current version of the Metaphor Ontology is available as Appendix 6.

Future work

The future work can be seen in several directions:

- finishing the Metaphor Ontology and related lexicons
- adding more languages to the metaphor modelling as well as to the topic and sentiment modelling experiments
- investing work in making the resources LOD
- injection of linguistic resources into large language models to improve their performance and to evaluate and understand their behaviour.

2.4.2 UC4.4.2 Use Case in Pharmacology

Coordinator: Dimitar Trajanov (University ss. Cyril and Methodius – Skopje)

Overview

The main objective of this use case was to survey the recent use of NLP in the field of pharmacology. As our work shows, NLP is a highly relevant information extraction and processing approach for pharmacology. It has been used extensively, from intelligent searches through thousands of medical documents to finding traces of adversarial drug interactions in social media. We split our coverage into five categories to survey modern NLP: methodology, commonly addressed tasks, relevant textual data, knowledge bases, and useful programming libraries. We split each of the five categories into appropriate subcategories, describe their main properties and ideas, and summarize them in a tabular form. The resulting survey presents a comprehensive overview of the area, useful to practitioners and interested observers.

Resources, methods, tools/technologies, and languages used

In this use case, we identified several sets of pharmacology-relevant datasets

Patient Data

- Collection of administrative claims. Size=43600000 Type=(NER, ADE, DDI) Link= https://www.ibm.com/products/marketscan-research-databases
- Data on patients hospitalized. Size=40000 Type=(Drug discovery, ADE, DDI) Link=https://mimic.mit.edu/
- A challenge dataset with 21 EHRs of cancer patients. Size=1089 Type=(NER, ADE) Link= https://bio-nlp.org/index.php/announcements
- Unstructured notes from the Research Patient Data. Size=505 Type=(ADE) Link= https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

Drug Usage Data

- Drug label database. Size=142981 Type=(NER, DDI, ADE) Link=https://dailymed.nlm.nih.gov/dailymed/index.cfm
- Database of drugs and drug products. Size=14665 Type=(ADE, pharmacovigilance, standardization, interactions) Link=https://go.drugbank.com/

Drug Structure Data

- Binding, functional, and ADMET data. Size=2.4 million Type=(ADE, pharmacovigilance, standardization, interaction) Link=https://www.ebi.ac.uk/chembl/
- Biomedical vocabularies. Size=2 million Type=(ADE) Link=http://umlsks.nlm.nih.gov
- Biologic macromolecules. Size=133920 Type=(ADE, pharmacovigilance, standardization, interaction) Link=http://www.rcsb.org/pdb/

- Biologic annotations. Size=1820 Type=(ADE) Link=https://biocreative.bioinformatics.udel.edu/news/corpora/chemprot-corpus-biocreative-vi/

Question Answering Data

- Collection of medical related pairs of questions and answers. Size=3048 Type=(QA) Link= https://github.com/curai/medical-question-pair-dataset
- Collection of COVID-19-related questions divided into 15 general categories and 207 specific question classes. Size=1690 Type=(QA) Link=https://paperswithcode.com/dataset/covid-q
- Collection of question–article–answer triplets taken from 85 different articles in CORD-19. Size=124 Type=(QA) Link=https://aclanthology.org/2020.nlpcovid19-acl.18/

General Pharmacological Data

- Wikipedia Online free encyclopedia. Size=15 billion Type=(ADE, DDI, drug discovery, NER) Link= https://en.wikipedia.org/
- Web engine for searching health articles. Size=30 million Type=(ADE, DDI, drug discovery, NER) Link= https://pubmed.ncbi.nlm.nih.gov/
- Scientific PubMed articles related with COVID-19. Size=255935 Type=(ADE, DDI, drug discovery, NER) Link=https://www.ncbi.nlm.nih.gov/research/coronavirus/
- Scientific papers relevant to COVID-19 research. Size=52000 Type=(ADE, DDI, drug discovery, NER) Link=https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge
- Dbpedia: Articles and structured data on e.g., drugs and diseases. Size=10000 Type=(ADE, pharmacovigilance, standardization, interactions) Link=https://www.dbpedia.org/
- Biologic and pharmacological bibliographic database. Size=32 million Type=(ADE, pharmacovigilance, standardization, interactions) Link= http://www.embase.com
- Clinical trials database. Size=329000 Type=(ADE, pharmacovigilance, standardization, interactions) Link=https://www.clinicaltrials.gov/

Knowledge graphs from the biomedical domain

- Bio2RDF. Unique Entities=1107871027 RDF Statements=11895348562
- HIFM. Unique Entities=3000 RDF Statements=21233
- LinkedDrugs. Unique Entities=248746 RDF Statements=99235032
- Covid-19-DS. Unique Entities=262954 RDF Statements=69434763
- KG-Covid-19. Unique Entities=574778 RDF Statements=24145556

Roadmap, workflow and milestones

The use case began with an in-depth review of NLP-related datasets, models, and libraries that are used in the pharmaceutical domain. Then, we created a systematic approach to summarize all the information and present it in a survey paper.

Deliverables

- A comprehensive overview of the current state in the area after the rapid developments that occurred in the past few years.
- Trajanov, Dimitar, Vangel Trajkovski, Makedonka Dimitrieva, Jovana Dobreva, Milos Jovanovik, Matej Klemen, Aleš Žagar, and Marko Robnik-Šikonja. "Review of Natural Language Processing in Pharmacology." Pharmacological Reviews 75, no. 4 (2023): 714-738.

Future work

We plan to update the review to include the latest advances in the NLP area, including Large Language Models.

3. Related activities

3.1 Collaboration

3.1.1 Collaboration within NexusLinguarum

Cross-WG collaboration has been one of the axes underpinning this CA, becoming visible not only through regular participation of different CA members in dedicated WG and UC meetings but also through joint publications and event organization. As stated in D4.2, 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), to help support these interactions and make the potential connections between WGs and their different tasks more explicit. The extended matrix played a key role in fostering further development in the subsequent GPs. Figures 4-6 represent the overall cross-WG interactions.

Figure 4 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 requirements and challenges identified by the various Use Cases. In addition, quality (T1.4), metadata, and versioning (T1.2) have also been addressed 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, discourse relations	applying LLOD for bilingual termbase data linking			
T1.4 Sources quality								
T1.5 Under-resourced languages								

Figure 4: *WG1-WG4 matrix*

Initial contacts were also established at the Skopje meeting between WG2 and WG4 (cf. Figure 5), mainly focusing on Knowledge extraction (T2.1) and on Terminology and Knowledge Management (T2.5), areas which constituted points of interest across practically all WG4 Use Cases and continued to strengthen during the following GPs.

	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 5: *WG2-WG4 matrix*

The preliminary contacts established between WG4 and WG3 at the Skopje meeting continued to intensify throughout the subsequent GPs, 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 6: *WG3-WG4 matrix*

In what concerns cross-WG cooperation, special emphasis was given, on the one hand, to the fact that several datasets across the UCs 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.

UC4.1.3, which started in March 2022, is an excellent example of cross-WG cooperation. While initially originating from Task 3.2. on Deep Learning and LLOD, it also evolved from discussions with members of WG 2, 3, and 4. Given that the number of participants was between 22 and 25, the interest and engagement from all these working groups was very high.

The exchange sparked many discussions about differences across languages, as well as language-specific translation issues, providing interesting and valuable insights also into the limitations of data modelling across aligned languages.

One of the key deliverables of the NexusLinguarum CA which also actively promoted cross-WG collaboration was the Massive Open Online Course (MOOC) on LLOD. Originating from WG3 (T3.5), the MOOC was coordinated by Slavko Žitnik, one of the WG4 task leaders (T4.1), and comprised a total of 16 lessons, developed and prepared by several Nexus members, including both WG4 leaders and the UC4.1.3 coordinator.

Solid cooperation was also fostered and quite visible across UCs. For instance, UC4.1.1 and UC4.2.2 coordinators worked jointly on the topic of Common Ground, Alignment and Lexical Prediction in English, Bulgarian and Polish Dialogues (TED corpus and Polish National Corpus data), giving a presentation entitled *Ground-sharing and discourse development prediction associated with the role of prosody in the interpretation of communicative connectives,* at the *Societas Linguistica Europea Congress 2022*, in Bucharest. Another example was in UC4.4.2, with close collaboration being established between the University in Skopje and the University in Ljubljana and involving not only two members of WG4's core team (Dimitar Trajanov and Marko Robnik-Šikonja) but also other Nexus members.

3.1.2 Collaboration outside of NexusLinguarum

WG4 members have also been active in fostering collaboration with initiatives outside the scope of Nexus. UC4.3.1, for instance, cooperated with the W3C Ontology-Lexicon community group in the ongoing discussion on developing a Terminology module in Ontolex, as well as on how to model terminological datasets and enhance their interoperability by applying LLOD technologies.

Several WG4 members have also been participating in other, more recent, COST Actions whose topics somewhat interconnect with those of Nexus, namely:

- CA19102 – Language In The Human-Machine Era (LITHME)
- CA19134 – Distributed Knowledge Graphs (DKG)
- CA21129 – What are Opinions? Integrating Theory and Methods for Automatically Analyzing Opinionated Communication (*OPINION*)
- CA21167 – Universality, diversity and idiosyncrasy in language technology (*UniDive*)
- CA22126 – European Network On Lexical Innovation (ENEOLI)

Contacts were also established with members of the Emotion Conceptual Networks (EmoCNet) project, from the University of Rijeka, Croatia, dedicated to the corpus-based graph analysis and representation of the syntactic-semantic network patterns of emotion expression in public, political discourse and popular culture. On November 24, 2022, Benedikt Perak, the Team Leader of the project, attended the 16th WG4 telco to present the project, particularly the CongraCNet application.

Within UC4.1.1, and per invitation of its coordinator, a research and lecturing visit by Milana Bolatbek, Lecturer at the Al-Farabi Kazakh National University, took place for one week, in October 2023, at the University of Applied Sciences in Konin, Poland. The collaboration aimed at extending an offensive language taxonomy to non-Indo-European languages (Kazakh). This cooperation also included a lecture to MA students on “Offensive messages in Kazakh social media”.

3.2 Exchange

The intense work carried out in WG4 was leveraged by a considerable number of exchange visits, i.e., Short-Term Scientific Missions (STSM), as well as Virtual Mobility Grants (VMGs), especially after the pandemic restrictions were lifted. In the Grant Periods being reported (November 2021 to April 2024), encompassing GPs 3 to 5, 9 STSMs were carried out, as summarised in Table 3.

GP	Grantee	Host	Title	Dates	UC in WG4
3	Anna Bączkowska	Centre for Translation Studies – Univ. Vienna (Austria)	Offensive language translation in film discourse and its integration with LLOD methods: OntoLex-FrAC and VarTrans modules	02/04/2022 to 28/04/2022	UC4.1.1
3	Ilan Kernerman	Institute for the Dutch language (Netherlands)	Linking lexicographic resources to language proficiency level applications	03/04/2022 to 12/04/2022	evolved into workshops at LDK 2023 and eLex 2023
3	Giedre Valunaite-Oleskeviciene	Centre of Linguistics of the Univ. of Porto (Portugal)	Researching integrated implicit and explicit offensive language taxonomy	05/07/2022 to 10/07/2022	UC4.1.1
3	Purificação Silvano	Mozaika Ltd (Bulgaria)	Towards a multilingual lexicon of discourse markers	10/09/2022 to 17/09/2022	UC4.2.2
4	Paola Marongiu	King's College London (UK)	Lexical semantic change in Latin: a use case on medical Latin	08/05/2023 to 29/05/2023	UC4.2.1
4	Lucía Pitarch	Centre for Translation Studies – Univ. Vienna (Austria)	Population of LLOD cloud with Deep learning approaches: Metaphor conceptualization and multilingual lexical relation acquisition	01/09/2023 to 01/10/2023	UC4.1.3
4	Barbara Lewandowska-Tomaszczyk	Masaryk University, Brno (Czech Republic)	Persuasion by emotion in social media – threat and stance identification for computer applications	03/09/2023 to 10/03/2023	UC4.1.1
4	Hugo Gonçalo Oliveira	Centre for Translation Studies – Univ. Vienna (Austria)	Relation Acquisition from Large Language Models (LLMs): Approaches and Evaluation	11/09/2023 to 30/09/2023	UC4.1.3
4	Aleksandra Tomaszewska	Centre of Linguistics of the Univ. of Porto (Portugal)	Multilingual Discourse Annotation Initiative (MDAI): A Polish-Portuguese Cooperation in Discourse Relations Annotation	04/09/2023 to 31/10/2023	UC4.2.2

Table 3. STSMs involving WG4 members (GP3 & GP4)

Furthermore, 3 VMGs were awarded in GP5 to WG4 core members Sara Carvalho, Slavko Žitnik, and Dagmar Gromann to support the preparation of modules for the previously mentioned MOOC on LLOD, i.e., Module 6a: Linked Data and Terminology, Module 3b: SPARQL, and Module 6b: Deep Learning and LLOD, respectively. As regards Slavko Žitnik, the VMG also aimed to support his coordination of the entire MOOC. Finally, an ITC Grant was assigned to Sara Košutar in GP3 (May 2022), supporting her participation in LREC 2022 to present a poster related to UC4.1.2 (cf. Section 2.1.2).

3.3 Events

During the period under analysis (GP3 to GP5), event organisation intensified within WG4, involving several of its members. Many of such events built upon previous Nexus initiatives, having more than one edition and spanning across Tasks and Use Cases, as can be seen below.

- Workshop on **Taxonomy and annotation of offensive language: implicitness**. Co-located with the NexusLinguarum Workshop Days. Jerusalem College of Technology, Israel. 23-24 May 2022. [UC4.1.1]
- Workshop on **Annotation scheme and evaluation: the case of OFFENSIVE language**. Co-located with the NexusLinguarum 4th Plenary Meeting in Vilnius and the LLODREAM Conference. Mykolas Romeris University, Lithuania. 21-22 September 2022. [UC4.1.1]
- Workshop on **Languages of Offence in Digital Perspectives**. Co-located with the International Conference Contacts & Contrasts. Languages in Cultural Perspectives: Practices, Discourses, Cognition. University of Applied Sciences in Konin, Poland. 27-29 March 2023. [UC4.1.1]
- **2nd Workshop DL4LD: Addressing Deep Learning, Relation Extraction, and Linguistic Data**. Co-located with the NexusLinguarum 4th Plenary Meeting in Vilnius and the LLODREAM Conference. Mykolas Romeris University, Lithuania. 22 September 2022. [UC4.1.3]
- **3rd Workshop DL4LD: Addressing Deep Learning, Relation Extraction, and Linguistic Data with a Case Study on The Bigger Analogy Test Set (BATS)**. Co-located with the 4th Conference on Language, Data, and Knowledge. University of Vienna, Austria. 13 September 2023. [UC4.1.3]
- **1st Workshop on Discourse studies and linguistic data science (DisLiDas)**. Co-located with the NexusLinguarum Workshop Days. Jerusalem College of Technology, Israel. 23-24 May 2022. [UC4.2.2]
- **2nd Workshop on Discourse studies and linguistic data science (DisLiDas)**. Co-located with the 4th Conference on Language, Data, and Knowledge. University of Vienna, Austria. 13 September 2023. [UC4.2.2]
- **Workshop on Terminology and LLOD in Life Sciences**. Co-located with the Terminology and Specialized Knowledge Representation Conference (TESK 2022). University of Rijeka, Croatia. 9-10 June 2023. [Task 4.3]

In addition to the events outlined above, both WG4 leaders were involved in the organisation of other, cross-WG initiatives in GP3 and GP4:

- **Workshop on Linking Lexicographic and Language Learning Resources (4LR)**. Co-located with the 4th Conference on Language, Data, and Knowledge. University of Vienna, Austria. 13 September 2023.

- **2nd Workshop on Sentiment Analysis & Linguistic Linked Data (SALLD-2)**. Co-located with the LREC 2022 Conference. Marseille, France. 24 June 2022.
- **3rd Workshop on Sentiment Analysis & Linguistic Linked Data (SALLD-3)**. Co-located with the 4th Conference on Language, Data, and Knowledge. University of Vienna, Austria. 13 September 2023.
- **Tutorial on Trends in Terminology Generation and Modelling (TermTrends)**. Co-located with the 23rd International Conference on Knowledge Engineering and Knowledge Management. Bozen-Bolzano, Italy. 26 September 2022.
- **Workshop on Terminology in the Era of Linguistic Data Science (TermTrends 2023)**. Co-located with the 4th Conference on Language, Data, and Knowledge. University of Vienna, Austria. 13 September 2023.
- **Workshop on Models and Best Practices for Terminology Representation in the Semantic Web (TermTrends24)**. Co-located with the 3rd International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT). Granada, Spain. 26 June 2024.

4. Concluding remarks and prospects

This final report outlines the progress within WG4 and its Use Cases. It focused especially on GP3 to GP5 (November 2021 – April 2024), given that a comprehensive description of the initial GPs had already been provided in the previous deliverables, in April and October 2021, respectively. Overall, significant work has been carried out in the nine UCs and across the four Tasks, with the development of new datasets and tools (close to 20 altogether) characterised by a strong multilingual focus. The close to 90 publications developed within the scope of WG4 (encompassing journal articles, papers in conference proceedings, as well as posters), along with the roughly 50 presentations in conferences and workshops, also attest to the massive amount of work carried out by the 138 members of the WG.

Intra-WG cooperation was fostered and supported by the complementary linguistic and computational backgrounds of our members, especially in devising LD-compliant solutions for resource creation and dissemination. The topics of SA and LLD, as well as knowledge organisation and extraction, emerged prominently across many UCs and were further explored throughout the CA. The Sentiment Analysis and Linguistic Linked Data (SALLD) workshop series, initiated within the scope of WG4, under the support of NexusLinguarum, illustrates the importance of such topics for this CA.

In addition, inter-WG cooperation was also successful. Stronger connections were established with WG1 in the first two GPs, mainly addressing the initial UC challenges regarding modeling and interlinking, but the engagement with other WG1 tasks, namely concerning quality and metadata, became increasingly visible. Collaboration with WG2 became more regular from GP3 onwards, especially given the importance of terminological data across the various UCs. The dataset developed within UC4.3.1, for instance, has provided relevant examples in the debate concerning the possible creation of a Terminology module in Ontolex. Moreover, WG2, WG3 and WG4 core team members started a workshop series on Terminology and Linked Data titled TermTrends, while also developing a module about the same topic for the MOOC on LLOD. The collaboration with WG3 gained traction via the new UC4.1.3 on Acquiring RDF Relations with Neural Language Models, managing to involve several dozen CA members and the subsequent dataset with 15 languages. Furthermore, multidimensional LD, particularly the challenges underlying the modeling of such data across space and time, was particularly relevant for the work carried out in UC4.2.1.

To help foster the aforementioned intra- and inter-WG cooperation, WG4 supported UC-related applications for STSMs and VMGs. Overall, 13 STSMs (9 in GP3 and GP4) and 3 VMGs were carried out throughout the Action, the latter related to the development of the MOOC. There was also one ITC grant assigned to one of the young researchers collaborating in the WG (namely in UC4.1.2).

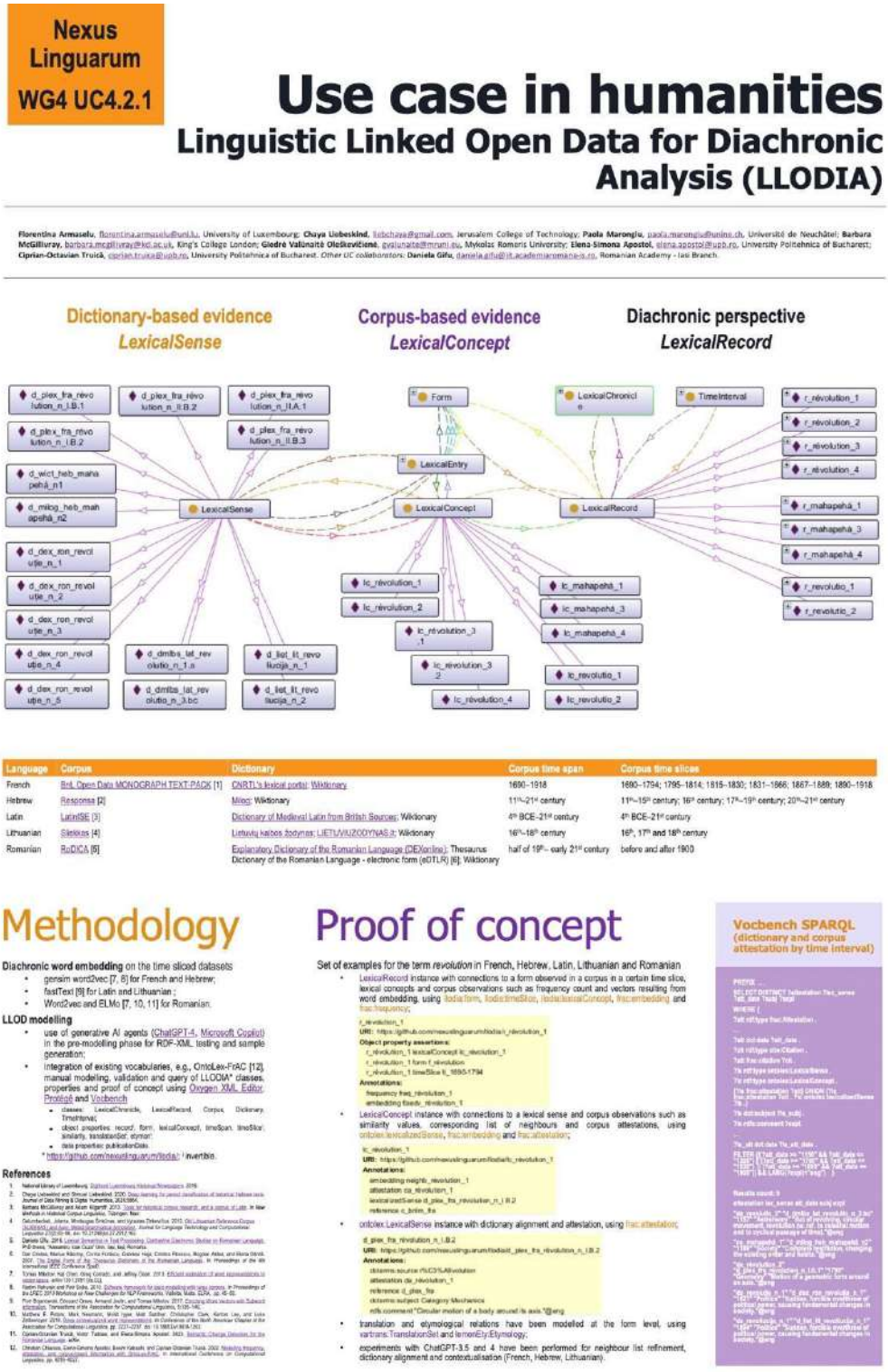
Collaboration with other COST Actions, along with other networks and projects (such as CLARIN, the European Language Grid, the W3C Ontology Lexica and BPMLOD Community Groups) was also paramount in initiating and/or strengthening contacts and partnerships.

Throughout the Action, WG4 leadership strongly encouraged event organisation and this, in turn, helped boost and showcase the work developed across the Tasks and UCs. Overall, 16 events (14 in GP3 through GP5), mainly workshops, were (co-)organised by a considerable number of WG4 members, mainly from the core team.

In conclusion, it is believed that WG4 managed to accomplish all its objectives, and has, in fact, exceeded them, largely due to the outstanding level of commitment of those more heavily involved in the WG activities. The strong connections established among many of the WG participants, substantiated in event organisation, joint publications and projects, represent an optimistic vision for the near future. For most of the WG members, there is the expectation that the vibrant and engaging community that Nexus helped to develop can continue to grow, thus contributing to current and relevant research on Linguistic Linked Data.

Appendices

Appendix 1: UC 4.2.1 [poster]



Language	Corpus	Dictionary	Corpus time span	Corpus time slices
French	BnL Open Data MONOGRAPH TEXT-PACK [1]	CNRTL's lexical portal; Wiktionary	1690–1918	1690–1794; 1795–1814; 1815–1830; 1831–1866; 1867–1889; 1890–1918
Hebrew	Responsa [2]	Milog; Wiktionary	11th–21st century	11th–15th century; 16th century; 17th–19th century; 20th–21st century
Latin	LatinISE [3]	Dictionary of Medieval Latin from British Sources; Wiktionary	4th BCE–21st century	4th BCE–21st century
Lithuanian	Sielokas [4]	Lietuvių kalbos žodynas; LIETUVIUŽODYNAS.lt; Wiktionary	16th–18th century	16th, 17th and 18th century
Romanian	RoDICA [5]	Explanatory Dictionary of the Romanian Language (DEXonline); Thesaurus Dictionary of the Romanian Language - electronic form (eDTLR) [6]; Wiktionary	half of 19th– early 21st century	before and after 1900

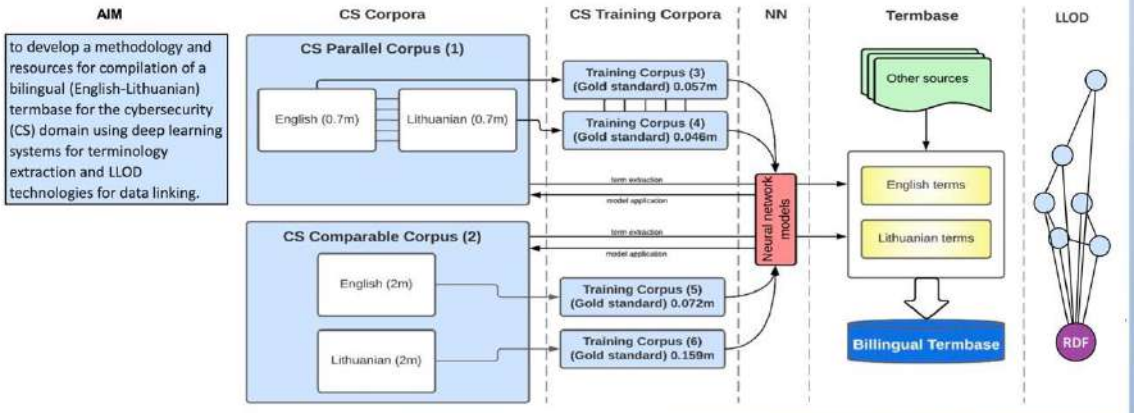
Appendix 2: UC 4.2.2 [poster]

Appendix 3: **UC 4.3.1 [poster]**

Use Case 4.3.1. Cybersecurity (1): Development of Language Resources

Sigita RACKEVIČIENĖ, Andrius UTKA

Funded by the European Union



COMPILED CORPORA

- **English-Lithuanian Parallel Cybersecurity Corpus** (TMX, linguistically annotated VERT)
- **English-Lithuanian Comparable Cybersecurity Corpus** (TXT, linguistically annotated VERT)

Both corpora are available at **CLARIN-LT repository**

CLARIN-LT

LITHUANIAN-ENGLISH CYBERSECURITY TERMBASE

234 CS concepts desginated by

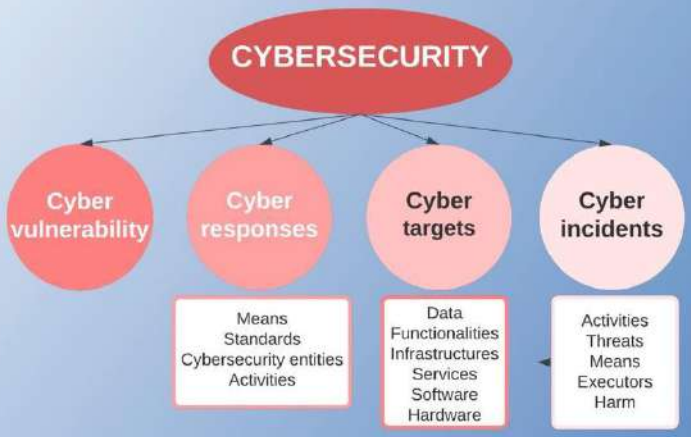
- 582 LT terms
- 427 EN terms (December, 2023)

- Publicly available for searching terms, their definitions, context examples, etc. on **terminologue.org/csterms**
- TBX file is freely available at CLARIN-LT repository.

EN-LT PARALLEL CORPUS (2010-2020)

The EU documents extracted from EUR-Lex and other EU institutional repositories:

- **Legally binding documents:** regulations, directives, decisions of the EU Parliament of Council;
- **Non-binding documents**: communications, reports and recommendations of the EU Commission, opinions of the EU Committees, briefing papers of the Court of Auditors.



EN-LT COMPARABLE CORPUS (2010-2021)

- **Legal documents:** CS strategies, laws, government resolutions, minister orders;
- **Administrative-informative texts:** reports and recommendations of CS centres, booklets, posters;
- **Academic publications:** scientific papers, books, theses, textbooks;
- **Mass and specialised media articles.**

Mykolas Romeris University

VYTAUTO DIDŽIOJO UNIVERSITETAS

Appendix 4: UC 4.3.2 [poster]

Automating the process of dictionary creation for sentiment analysis in finance

Authors

Maryan Rizinski (1,2), Hristijan Peshov (2), Kostadin Mishev (2) , Milos Jovanovik (2), Dimitar Trajanov (1,2)

Affiliations

1) Department of Computer Science, Metropolitan College, Boston University, Boston, MA 02215, USA

2) Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, 1000, Skopje, North Macedonia

We introduce a novel methodology named eXplainable Lexicons (XLex), which offers a solution to automating the creation of sentiment dictionaries in finance. Traditionally, sentiment analysis in finance has relied on manually annotated lexicons, which demands significant effort from human experts to develop, maintain, and update these specialized resources. Despite the simplicity and speed of lexicon-based approaches, they fall short compared to deep learning methods, like transformer models, known for their superior performance in natural language processing tasks. However, transformer models require extensive data and computational resources and have longer prediction times, limiting their practicality in real-time or resource- constrained environments.

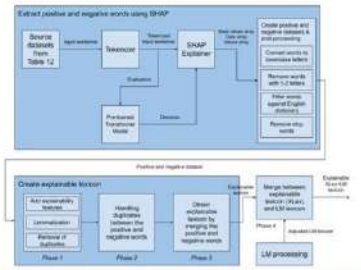
1 Introduction

Lexicon-based methods are simple to implement and fast to operate on textual data, but they require considerable manual annotation efforts to create, maintain, and update the lexicons. These methods are also considered inferior to the deep learning-based approaches, such as transformer models, which have become dominant in various natural language processing (NLP) tasks due to their remarkable performance. However, their efficacy comes at a cost: these models require extensive data and computational resources for both training and testing.

We present a novel XLex methodology that leverages NLP transformer models and SHAP explainability to automatically enhance the vocabulary coverage of the Loughran-McDonald (LM) lexicon in sentiment analysis scenarios for financial applications. Our results demonstrate that standard domain-specific lexicons, such as the LM lexicon, can be expanded in an explainable way with new words without the need for laborious annotation involvement of human experts, a process that is both expensive and time-consuming.

2 Methodology

The architecture of the data processing pipeline for generating the explainable lexicon (XLex).



3 Mathematical model

The sentiment analysis model calculates the sentiment of every sentence. The optimal parameters of the model are determined using a grid search procedure.

$$sentence = \{w_1, w_2, \ldots, w_n\} \quad (1)$$

$$v_c^{pos}(w_i) = \sum_{i=1}^{n} v^{pos}(x_i) \quad (2)$$

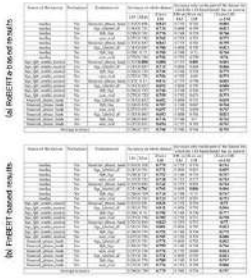
$$v_{sent}(w_i) = c_{dp} * v_c^{sl}(w_i) + c_{dn} * v_c^{sl\_opp}(w_i) + + c_{lmp} * v_c^{lm}(w_i) + c_{lmn} * v_c^{lm\_opp}(w_i) \quad (3)$$

$$v_{sent}(sentence) = \sum_{i=1}^{n} v_{sent}(w_i) \quad (4)$$

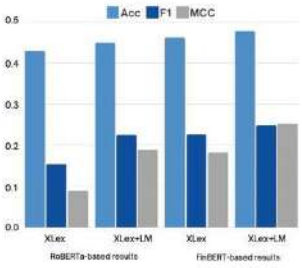
$$s_{pol}(sentence) = \begin{cases} positive : v_{sent}(sentence) > 0 \\ negative : v_{sent}(sentence) < 0 \\ neutral : otherwise \end{cases} \quad (5)$$

4 Results

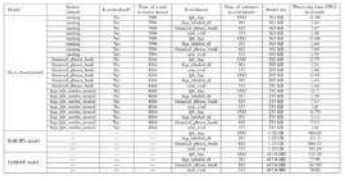
The accuracy of the XLex methodology for models based on the RoBERTa and FinBERT transformers.



Average improvements of XLex and XLex+LM over LM in terms of accuracy, F1, and MCC.



Comparison of the XLex-based model with the RoBERTa and FinBERT transformers-based models in terms of model speed and size. Tested on a server with Intel Xeon CPU at 2.20GHz and 12GB RAM.



XLex achieved a speed-up:
- **87x** faster compared with RoBERTa
- **21x** faster compared with FinBERT

XLex Model size:
- **4400x** smaller compared with RoBERTa
- **1340x** smaller compared with FinBERT

5 Conclusion

- The novel XLex methodology leads to significant improvements in sentiment analysis.
- Our experiments reveal that XLex achieves higher accuracy and larger vocabulary coverage, directly addressing the limitations of standard, manually annotated lexicons.
- Additionally, this methodology is substantially more efficient in terms of model speed and size when compared to transformer models. omains beyond finance.
- XLex exhibits inherent interpretability, a crucial feature that facilitates a deeper understanding and insight into sentiment analysis results, making it a valuable asset for financial decision-making.
- The proposed methodology is general and adaptable, offering opportunities for future research to explore its application across other domains beyond finance.

"St. Cyril and Methodius" University in Skopje FACULTY OF COMPUTER SCIENCE AND ENGINEERING

BOSTON UNIVERSITY

Funded by the European Union

Appendix 5: **UC 4.4.1 [poster]**

Use Case 4.4.1 in Public Health

Marko Robnik-Šikonja (University of Ljubljana, Faculty of Computer and Information Science)
Petya Osenova (Bulgarian Academy of Science, IICT)
Ana Ostroški Anić (Institute for the Croatian Language)

Focus

Disease prevention and Quality of life

Main tasks

Discover tendencies in news media, social media, etc.
Cross-lingual setting

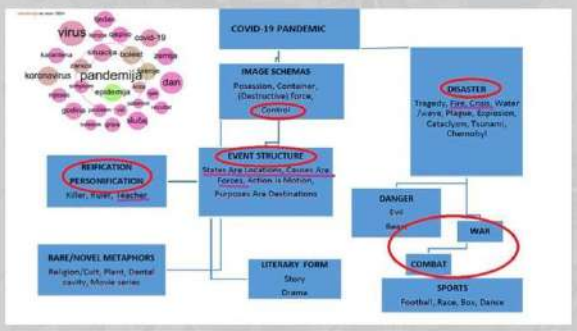
Multilingual metaphors in COVID-19 pandemic

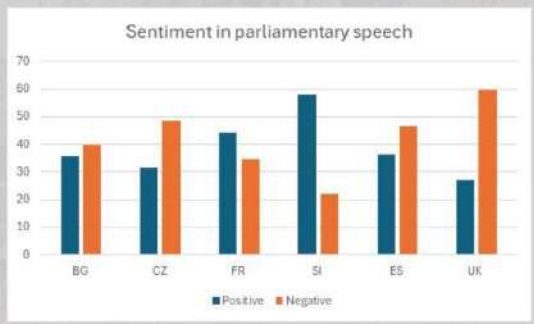
metaphor varieties related to the pandemic situation in Bulgarian, Croatian, Lithuanian, and Slovenian

initial modeling of **metaphor frames into an ontology** and its linking to relevant lexica

Multilingual parliamentary speech analysis during COVID-19 pandemic

- **six parliaments:** Bulgarian, Czech, French, Slovenian, Spanish, UK
- **cross-lingual topic comparison**
- **sentiment, emotions**
- **differences in language** based on speakers' gender, age, and political orientation





Kristina Despot, Liudmila Mockienė, Petya Osenova and Ana Ostroški Anić: Covid-19 and health related metaphors in Bulgarian, Croatian and Lithuanian ParlaMint corpora. Presentation at: *Terminology and Specialized Knowledge Representation: New Perspectives on User Needs* (TESK 2022).

Kristian Miok, Encarnacion Hidalgo-Tenorio, Petya Osenova, Miguel-Angel Benitez-Castro and Marko Robnik-Šikonja. Multi-aspect Multilingual and Cross-lingual Parliamentary Speech Analysis. *Intelligent Data Analysis*. 2024.

Further work

- Finalizing the Metaphor Ontology and related lexicons
- Adding more languages to the metaphor modeling
- Making the resources LOD
- Adding more languages to the topic, sentiment, emotions etc. modeling
- Adding LRs into LLMs to improve their performance and to evaluate and understand their behavior.

Funded by the European Union

Appendix 6: Encoding of the Metaphor Ontology

Pref nexuso = http::/www.nexus-ling.eu/metaphorOnto.

%%%%%%%%%

nexuso:CovidFrame rdf:type owl:Class .

nexuso:Disaster rdf:type owl:Class .

nexuso:Disaster rdf:subClass owl:CovidFrame .

nexuso:Water rdf:subClass nexuso:Disaster .

nexuso:WaterWave rdf:subClass nexuso:Water .

nexuso:Tsunami rdf:subClass nexuso:WaterWave .

nexuso:Fire rdf:subClass nexuso:Disaster .

nexuso:Tragedy rdf:subClass nexuso:Disaster .

nexuso:Fire rdf:subClass nexuso:Disaster .

nexuso:Crisis rdf:subClass nexuso:Disaster .

nexuso:Plague rdf:subClass nexuso:Disaster .

nexuso:Explosion rdf:subClass nexuso:Disaster .

nexuso:Cataclysm rdf:subClass nexuso:Disaster .

nexuso:Chernobyl rdf:subClass nexuso:Disaster .

nexuso:Reification rdf:type owl:Class .

nexuso:Reification rdf:subClass owl:CovidFrame .

%%%%%%%%%

nexuso:Personification rdf:type owl:Class .

nexuso:Personification rdf:subClass owl:CovidFrame .

nexuso:Killer rdf:subClass nexuso:Personification .

nexuso:Ruler rdf:subClass nexuso:Personification .

nexuso:Teacher rdf:subClass nexuso:Personification .

nexuso:Partner rdf:subClass nexuso:Personification .

%%%%%%%%%

nexuso:ImageSchemas rdf:type owl:Class .

nexuso:ImageSchemas rdf:subClass owl:CovidFrame .

nexuso:Posession rdf:subClass nexuso:ImageSchemas .

nexuso:Container rdf:subClass nexuso:ImageSchemas .

nexuso:DestructiveForce rdf:subClass nexuso:ImageSchemas .

nexuso:Control rdf:subClass nexuso:ImageSchemas .

%%%%%%%%%

nexuso:EventStructure rdf:type owl:Class .

nexuso:EventStructure rdf:subClass owl:CovidFrame .

nexuso:States rdf:subClass nexuso:EventStructure .

nexuso:AreLocations rdf:subClass nexuso:EventStructure .

nexuso:CausesAreForces rdf:subClass nexuso:EventStructure .

nexuso:ActionIsMotion rdf:subClass nexuso:EventStructure .

nexuso:Purposes rdf:subClass nexuso:EventStructure .

nexuso:AreDestinations rdf:subClass nexuso:EventStructure .

%%%%%%%%%

nexuso:Danger rdf:type owl:Class .

nexuso:Danger rdf:subClass owl:CovidFrame .

nexuso:Evil rdf:subClass nexuso:Danger .

nexuso:Beast rdf:subClass nexuso:Danger .

%%%%%%%%%

nexuso:LiteraryForm rdf:type owl:Class .

nexuso:LiteraryForm rdf:subClass owl:CovidFrame .

nexuso:Story rdf:subClass nexuso:LiteraryForm .

nexuso:Drama rdf:subClass nexuso:LiteraryForm .

%%%%%%%%%

nexuso:Sports rdf:type owl:Class .

nexuso:Sports rdf:subClass owl:CovidFrame .

nexuso:Football rdf:subClass nexuso:Sports .

nexuso:Race rdf:subClass nexuso:Sports .

nexuso:Box rdf:subClass nexuso:Sports .

nexuso:Dance rdf:subClass nexuso:Sports .

%%%%%%%%%

nexuso:RareOrNovelMetaphors rdf:type owl:Class .

nexuso:RareOrNovelMetaphors rdf:subClass owl:CovidFrame .

nexuso:Religion rdf:subClass nexuso:Sports .

nexuso:Cult rdf:subClass nexuso:Sports .

nexuso:Plant rdf:subClass nexuso:Sports .

nexuso:DentalCavity rdf:subClass nexuso:Sports .

nexuso:MovieSeries rdf:subClass nexuso:Sports .

%%%%%%%%%

nexuso:War rdf:type owl:Class .

nexuso:War rdf:subClass owl:CovidFrame .

%%%%%%%%%

nexuso:Combat rdf:type owl:Class .

nexuso:Combat rdf:subClass owl:CovidFrame .

%%% THREAT

nexuso:Threat rdf:type owl:Class .

nexuso:Threat rdf:subClass owl:CovidFrame .

%%%% The Following are copied from Lemon, but modified

nexuso:LexicalEntry

a rdfs:Class, owl:Class ;

rdfs:comment "An entry in the lexicon. This may be any morpheme, word, compound, phrase or clause that is included in the lexicon"@en ;

rdfs:comment "It is copied from Lemon, but modified"@en ;

rdfs:label "Entrada léxica"@es, "Entrée lexicale"@fr, "Lexical entry"@en, "Lexikaal item"@nl, "Lexikoneintrag"@de ;

rdfs:subClassOf :HasLanguage, :HasPattern, :LemonElement, [

a owl:Restriction ;

owl:minCardinality "1"^^xsd:nonNegativeInteger ;

owl:onProperty nexuso:lexicalForm

] .

nexuso:lexicalForm

a rdf:Property, owl:InverseFunctionalProperty, owl:ObjectProperty ;

rdfs:comment "Denotes a written representation of a lexical entry"@en ;

rdfs:domain nexuso:LexicalEntry ;

rdfs:label "Forma léxica"@es, "Forme lexicale"@fr, "Lexical form"@en, "Lexikaal vorm"@nl, "Lexikalische Form"@de ;

```
        rdfs:range nexuso:Form .
nexuso:Form
        a rdfs:Class, owl:Class ;
        rdfs:comment "A given written or spoken realisation of a lexical entry"@en ;
        rdfs:label "Form"@de, "Form"@en, "Forma"@es, "Forme"@fr, "Vorm"@nl ;
        rdfs:subClassOf nexuso:LemonElement, [
        a owl:Restriction ;
        owl:minCardinality "1"^^xsd:nonNegativeInteger ;
        owl:onProperty nexuso:representation
        ] .
nexuso:representation
        a rdf:Property, owl:DatatypeProperty ;
        rdfs:comment "A realisation of a given form"@en ;
        rdfs:domain nexuso:Form ;
        rdfs:label "Darstellung"@de, "Representación"@es, "Representation"@en,
"Représentation"@fr, "Voorstelling"@nl ;
        rdfs:range xsd:string .
nexuso:writtenRep
        a rdf:Property, owl:DatatypeProperty ;
        rdfs:comment "Gives the written representation of a given form"@en ;
        rdfs:domain nexuso:Form ;
        rdfs:label "Representación escrita"@es, "Représentation écrite"@fr, "Schriftelijke
voorstelling"@nl, "Schriftliche Darstellung"@de, "Written representation"@en ;
        rdfs:range xsd:string ;
```

rdfs:subPropertyOf nexuso:representation .

%%%%

nexuso:frame

a rdf:Property, owl:InverseFunctionalProperty, owl:ObjectProperty ;

rdfs:comment "Link a lexical entry with a methafor frame"@en ;

rdfs:domain nexuso:LexicalEntry ;

rdfs:label "Forma léxica"@es, "Forme lexicale"@fr, "Lexical form"@en, "Lexikaal vorm"@nl, "Lexikalische Form"@de ;

rdfs:range nexuso:CovidFrame .