An overview of NexusLinguarum use cases: Current status and challenges

Sara Carvalho and Ilan Kernerman

Introduction

Working Group 4 (WG4) of the NexusLinguarum COST Action – European network for Web-centred linguistic data science (CA18209) – is dedicated to applying and validating the Action's methodologies and technologies, as well as to eliciting the requirements for their implementation.

With more than a hundred members from nearly 40 countries, WG4, entitled *Use Cases and Applications*, is the biggest working group in NexusLinguarum. The participants have expertise in various areas, thereby fostering interdisciplinary collaboration. WG4 includes four Tasks, each led by two persons with backgrounds in Linguistics and Computer Science, respectively, and including two Use Cases (UCs).

This article outlines WG4's Tasks and Use Cases, their objectives, requirements, methodologies, resources, milestones and expected deliverables. In addition, it describes the cooperation with the other WGs of NexusLinguarum.

Working Group 4

WG4

Sara Carvalho is leader of NexusLinguarum WG4. She is an Assistant Professor at the University of Aveiro and a member of ISO TC37 (WG4/SC4 – Lexical resources). Her research interests include medical terminology and its interconnection with ontologies, health literacy, as well as specialized lexicography.

Ilan Kernerman is co-leader of NexusLinguarum WG4. He is CEO of K Dictionaries – Lexicala. His main interests include the intersection of multilingual linguistic data and knowledge systems.

Tasks and Use Cases

Task 4.1. Use Cases in Linguistics

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

More specific tasks will be accomplished within the particular use cases that are described in detail. During the first year of the COST Action (CA), two specific Use Cases have been shaped and the activities within those have been determined: Media and Social Media, on the one hand, and Language Acquisition. There is a possibility of adding other use cases in linguistics in the following CA years.

UC 4.1.1. Use Case in Media and Social Media

UC 4.1.2. Use Case on Language Acquisition

T4.1. leader, linguistics

Kristina S. Despot is a Senior Research Fellow and Vice Director at the Institute for the Croatian Language and Linguistics. Her main research interest is cognitive linguistics, primarily figurative language. She has published five books and more than fifty research papers, and received a Fulbright Postdoctoral Award.

T4.1. leader, computational

Slavko Žitnik is Assistant Professor at the University of Ljubljana, Faculty for Computer and Information Science. His research interests are mainly in natural language processing related to semantic representations, i.e. knowledge extraction, knowledge bases and Semantic Web. He is engaged in multiple research and industry-related projects.

UC 4.1.1. Use Case in Media and Social Media

Overview. The principal aim of this use case is building 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 available ways to create an infrastructure regarding abusive data sharing. The UC team aims to modify and enrich the existing sentiment/emotion annotation tagsets and make an attempt to implement them into samples of the languages analysed. More specifically, this UC focuses on the development of abusive language event representation and scales, based on the typology and severity of offensive content (see Likert scales – severity scales of 5) in terms of multiple classifier tagsets. The tasks cover 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 (someone). 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. Small (social) media samples of relevant languages will be exemplified and analysed, in addition to their annotation and offensive content extraction.

UC4.1.1. coordinator

Barbara Lewandowska-Tomaszczyk is full Professor Dr habil. in Linguistics at the Department of Language and Communication of the State University of Applied Sciences in Konin (Poland). Her research focuses on cognitive semantics and pragmatics of language contrasts, corpus linguistics and their applications in translation, media studies, and incivility in online discourse.

Resources

- Big data: national language corpora, media and social media repositories, platforms; CLARIN
- Hate speech datasets: hatespeechdata.com (Derczynski & Vidgen, 2020)
- Samplers: small corpora of social media such as NLTK (Natural Language Toolkit), small collection of web texts, parts of EUROPARL
- Small datasets of languages represented in the use case (see 'Languages' below)

Methods

- 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
- Enrichment of explicit and implicit language tagsets towards abusive language extraction

Tools

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

The traditional methods (Naive Bayes, Support Vector Machines, Logistic Regression) will be useful for explicit abusive language, while contextual Deep Learning models based (e.g. ELMo) on transformer architectures, such as BERT, GPT, ELECTRA, will be 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.

Requirements

- Linguistic discourse analysis competence
- Linguistic competence to provide coding of speech samples in several languages
- Identification and familiarity with existing hate speech databases
- Identification and familiarity with existing hate speech tagset systems
- Abusive language and Sentiment Analysis extraction systems
- Tagging systems application
  - T1.1 support on corpus modelling

Languages

English, Croatian, Hebrew, Lithuanian, Montenegrin, Polish, Slovene

Roadmap

- Survey/selection of corpora
- An online workshop was held to discuss the computational aspects involved in each of the planned tasks (end of September 2020)
- Development of incivility/abuse identification scales (explicit, implicit)
- Identification of (multiple) tagset annotation tools
- Application (and enrichment) of tagset tools into English, Croatian, Hebrew, Lithuanian, Montenegrin, Polish, Slovene

Strategy

- The main aim of this use case is 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 will be the identification of abusive language corpora, as well as their annotation and extraction tools;
- The main strategy will cover the development of richer abuse event identification structure and identification scales;
- The main outcome will involve proposals concerning, on the one hand, a more detailed description of abusive language event structure and, on the other hand, relevant abusive language scales developed in the use case.

Tasks

T1. Description of details of the use case objectives and implementation: (abuse, implicitness/explicitness, emotions/sentiments, hate speech) – select 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; TYPES of abuse/accompanying EMOTIONS (Sentiment analysis+)

T5.a. Manual tagging of the selected data based on new decisions and scales in English and other languages

T6. Survey of automatic annotation tools and implementation of baseline models

T7. Abusive language tagset enrichment proposals

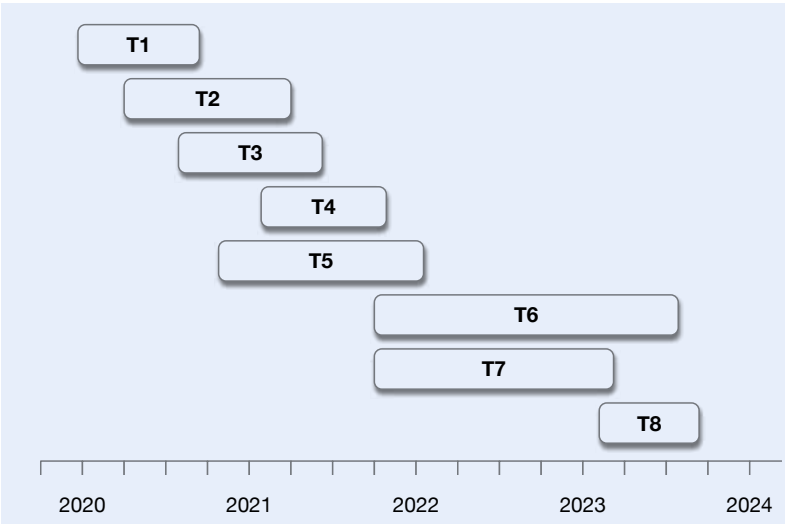
T8. Survey of LLOD infrastructure relevant to the task topic (ontologies of opinion mining, etc). Infrastructure proposals of abusive hate speech data sharing

Organisational task

An international conference on Explicit and Implicit Abuse is planned by UC4.1.1. members, including a workshop on *Social Media: analysis, extraction, LLOD*, for the Winter/Spring of 2022. The focus will be on various approaches to the theme with the aim to attract researchers from the other WGs, as well as other scholars from linguistics, media, psychology, and computer science. The workshop will present the work and results of the UC team.

Workflow

Duration: 01.06.2020 – 1.10.2023



Methodology

- Linguistic
  - Review and collection of the appropriate data
  - Definitions of abusive language categories and terminology, coding for the abusive language texts (scale)
  - Definition of tagging guidelines (for explicit and implicit examples) - level of annotation, figurative/not figurative, etc.
  - Preparation of a dataset for the computational models.
- Computational
  - Review of existing tools for abusive language identification
  - Work on the explicit abusive language detection and identification along with linked data representation of results
  - Enrichment of extracted linked data with existing automatically generated knowledge bases
  - Implementation of models for implicit abusive language detection

Deliverables

D1. Use case description and the identification of objective details (M12)

D2. Survey and acquisition of English hate speech corpus (M14)

D3. Acquisition of respective abusive language corpora (M24)

D4. Development of abusive language event representation and scales (explicit abuse) (M30)

D5. Development of abusive language event representation and scales (implicit abuse) (M36)

D6. Implementation of the enriched tagsets into samples of the languages analysed (M40)

D7. Survey of LLOD abusive tagset systems (M42)

D8. Final report and tagset enrichment proposal (M46)

Milestones

MS1. Acquisition of English hate speech corpus

MS2. Acquisition of relevant abusive language corpora

MS3. Proposals of explicit language abusive event structure description

MS4. Proposals of implicit language abusive event structure description

MS5. Development of abusive language event scales

Collaboration and Exchange

- UC coordination and WG4 communication channels
- Nexus WGs and WG4 UCs and Tasks (WG1 T1.1. (resources), WG4 T4.2. – Humanities and Social Sciences, WG4 UC4.3.1. Use Case in Cybersecurity, and others)
- Short Term Scientific Missions (STSMs)
- Other (beyond Nexus, if appropriate): CLARIN, TRAC, LREC

Dissemination

- Reports
- Meetings, workshops and activities
- An international conference on *Explicit and Implicit Abuse* is planned for spring 2022 and a workshop on Social Media: analysis, extraction, LLOD will be proposed within the scope of that conference
- Conferences: LREC, COLING, TRAC, Discourse and Semantics, ACL. EACL, LDK, Hate Speech conferences
- Publications – joint and individual – conference proceedings and journal publications

UC 4.1.2. Use Case on Language Acquisition

Overview. The aim of this use case is to promote the usage of web-based technologies for language acquisition, and to develop resources for that.

A language sample (written or spoken text produced an the individual, usually as a result of some language task like telling a story or writing an essay) provides information about first and second language acquisition or proficiency, i.e., can be used to assess the language of an individual speaker. Language sample analysis can be used by teachers of a second language, speech and language pathologists, elementary school teachers, employers in certain fields and so on. However, it has mostly been used within the fields of first and second language acquisition, that is, by speech and language pathologists and teachers of a second language. In both fields, same or similar measures have been used, but for the first language acquisition, language samples are usually spoken, whereas for the second language acquisition, they are usually written. This type of analysis is often used in some countries, but in many countries, scientists and professionals are unaware of its benefits.

A number of measures have been introduced in different domains (e.g. measures of productivity, measures of lexical diversity; overview of some: MacWhinney 2020). However, users often find the transcription and calculation of measures time-consuming (Pavelko et al. 2016). During the last decades of the 20th century, computer programs were developed to assist language sample analysis (overview: Pezold et al. 2020). Transcription, coding and analysis are not user-friendly in those programs, so they are more often used in the scientific community than by professionals. Lately, web-based programs for different aspects of analysis have been introduced, mainly developed within the scientific community (thus being open source), but still much more user-friendly than previous programs. Web-based programs usually concentrate on one domain. For example, the Gramulator tool (McCarthy et al. 2012) calculates different measures of lexical diversity. Coh-Metrix (Graesser et al. 2004) is more elaborate and includes several domains, all relevant for discourse analysis. Measures are based on basic calculations (e.g. type-token ratio, number of different words, mean length of a sentence), but there are also advanced measures based on language technologies. For example, a web-based application might include annotation of morphological and syntactic features, recognition of connectives or similar. Such annotation enadles the implementation of measures such as lemma-token ratio, lexical density (content words/number of words) or similar.

UC4.1.2. coordinator

Gordana Hržica is a linguist at the Department of Speech and Language Pathology of the University of Zagreb. Her scientific interests are language acquisition, bilingualism, language pathology and language processing. In prior studies she used a variety of methods, but mostly spoken language corpora and spoken language samples analysis.

Web services have been developed and are mostly used for English. Coh-Metrix has been adapted to other languages (Spanish, Portuguese, German), but not for the full range of measures and, as far as it is known, such adaptations are not publicly available.

There is, therefore, great potential in:

1. **Using existing language technologies to develop such tools for other languages**

Many languages already have technologies that can be used to annotate text in order to calculate a great range of measures, but possibly also to introduce new measures.

2. **Introduce new measures (e.g. based on linked data)**

Connecting with other language sources might allow advanced analyses. For example, data about the frequency of individual words or about the frequency of semantic structures can show us how frequent language elements used in the language sample are, which is the basis for calculating sophistication measures (Kyle et al. 2018). Other things that we are currently unaware of might be explored (e.g. using data from online dictionaries of different databases like those of metaphors of collocations).

3. **Promoting the usage of speech-sample analysis in different fields such as regular education**

Measures for analysis that have been developed and validated, such as measures of productivity and lexical diversity, implement basic calculations (e.g. type-token ratio, number of different words, mean length of a sentence).

However, there are also advanced measures based on language technologies. For example, a web-based application might include annotation of morphologic and syntactic features, and that would enable the implementation of measures like lemma-token ratio and syntactic density (percentage of subordinate clauses).

Resources

All languages and dialects can provide language samples to be analysed on a basic level. However, only some languages have sufficiently developed language technologies (e.g. morphological and syntactic taggers) for the application of advanced measures.

Methods

Measures applied for language sample analysis can be grouped into measures of: (1) productivity, (2) lexical richness, (3) syntactic complexity and (4) cohesion.

Tools

There are some existing computer programs used by the research community that are not user-friendly (CLAN – Computerized Language Analysis, SALT - Systematic Analysis of Language Transcripts). As mentioned earlier, some Web services have been developed and are mostly used for English (e.g. Gramulator), while Coh-Metrix has been developed for English, but also adapted to other languages (Spanish, Portuguese, German).

Requirements

- Linguistic competence to provide coding of speech samples in several languages
- Knowledge of the existing data sources
- Knowledge of the existing language technologies for different languages
- Knowledge about language acquisition measures (productivity, vocabulary diversity, syntax, discourse)

WG1 support for creating specific vocabularies (e.g. connectives, discourse markers, metaphoric usage of language), including from the contribution of T1.1 in developing best practices for defining specific usage.

WG3 support for establishing links between developed tool(s) and other corpora in order to retrieve data on frequency and collocations, needed to implement additional measures of language acquisition

Strategy

Our strategy is to gather information about the available general and language-specific tools for language sample analysis and to gain an overview of the methods of text analysis used in individual countries. This will help us develop strategies for the promotion of language sample analysis. During this period, we will outline a potential roadmap to the development of a web-based tool for language sample analysis in language acquisition.

Tasks

T1. Researching available language sample tools

T2. Researching available language technologies for participating languages

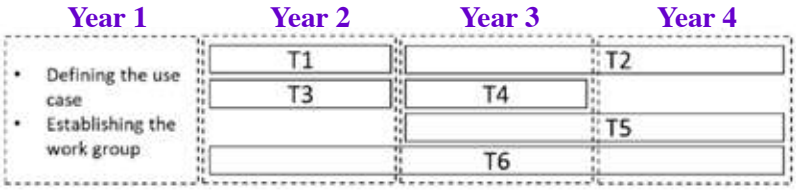
T3. Developing a survey for collecting information about the language sample analysis in individual countries

T4. Collecting information about the language sample analysis in individual countries

T5. Developing strategies for the promotion of language sample analysis

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

Workflow



Methodology

extensive literature research, scientific networking, online survey

Deliverables

D1. WG4 use case description for 4.1.2.

D2. NexusLinguarum use case template for 4.1.2.

D3. Overview of available language sampling tools

D4. Overview of language technologies for participating languages

D5. Survey for collecting information about the language sample analysis available online

D6. Participation of at least 50 researchers and/or practitioners in the survey

D7. Results of the survey analysed and presented

D7. An update of available language technologies for participating languages

D8. Web application for at least one of the participating languages developed

D9. Roadmaps for the development of web application for at least two participating languages

Milestones

M1. NexusLinguarum use case template for 4.1.2.

M2. Results of the survey analysed and presented

M3. Web application for at least one of the participating languages developed

Collaboration and Exchange

- UC coordination and WG4 communication channels
- WG4 UCs and Tasks
- Nexus WGs
- Short Term Scientific Missions (STSMs)
- Other: speech and language pathologists, language teachers, computational linguists

Dissemination

- Reports
- Meetings, Workshops, Conferences
- Publications

Task 4.2. Use Cases in Humanities and Social Sciences

Overview. This task focuses on how linguistic data science can deeply influence studies in the humanities, allowing us to trace the history of the peoples of the world, understand literature in new ways or predict and analyse social trends. It will also contribute to the social sciences by investigating the use and development of language processing tools that facilitate the usage of survey data archives.

As a use case in humanities, the task focuses on the evolution of parallel concepts in different languages, by establishing a set of guidelines for the construction of a comparative framework based on multilingual ontologies to represent semantic change through LLOD and Semantic Web technologies (e.g. ontolex-lemon, rdf, owl-time).

As a use case in social sciences, the task studies the ways in which survey data can be integrated, linked, processed and made accessible using LLOD methods. Such tools include data anonymization tools, semantic search, semantic data integration and relations detection.

UC 4.2.1. Use Case in Humanities

UC 4.2.2. Use Case in Social Sciences

T4.2. leader, linguistics (from June 2021)

Mietta Lennes is Project Planning Officer for the Language Bank of Finland. She supports students and researchers in language data management and teaches corpus linguistics and speech analysis. Her research interests are in the development of acoustic-phonetic analysis methods for conversational speech.

T4.2. leader, computational

Jouni Tuominen is a staff scientist at the Department of Computer Science, Aalto University, and a research coordinator at Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki. His research interests include ontology repositories and services, linked data publishing methods, ontology models for legacy data, and tooling for digital humanities.

UC 4.2.1. Use Case in Humanities

Overview. The use case focuses on the evolution of parallel concepts in different languages and Humanities fields (history, literature, philosophy, religion, etc). The methodology includes various textual collections and resources from corpus linguistics, word embedding and LLOD. This type of enquiry may provide evidence of changing contexts in which words pertaining to the target semantic fields appeared in different eras, enabling diachronic analysis and historical interpretation. The outcome will comprise a set of guidelines for constructing a comparative framework and a sample of multilingual ontologies to represent semantic change through LLOD and Semantic Web technologies.

UC4.2.1. coordinator

Florentina Armaselu is a research scientist at the Luxembourg Centre for Contemporary and Digital History of the University of Luxembourg. With a background in computer science, comparative literature and linguistics, her research interests focus on areas such as computational text analysis and text interpretation, text encoding, human-computer interaction, Semantic Web and digital humanities.

Resources.

- Historical textual corpora available in digital format (TXT, XML) and various domains of the Humanities (literature, philosophy, religion, history, etc): LatinISE (2nd century B. C. - 21st century A. D.) (McGillivray and Kilgarriff 2013); Diorisis (7th century BC - 5th century AD) (McGillivray et al. 2019; Vatri and McGillivray 2018); Responsa (11th century until now) (Liebeskind and Liebeskind 2020); the National Library of Luxembourg (BnL) Open Data collection (1841-1878, newspapers; 1690-1918, monographs) (Ehrmann et al. 2020); Sliekkas (16th to 18th century) (Gelumbeckaite et al. 2012).
- Lexicons and dictionaries, especially historical and etymological dictionaries from which information can be extracted (Khan 2020).

Methods

- Theoretical modelling of semantic change (Betti and Van den Berg 2014; Fokkens et al. 2016; Geeraerts 2010; Kuukkanen 2008; Wang et al. 2011).
- Expressing semantic change through LLOD formalisms (Khan 2018; Romary et al. 2019; Welty et al. 2006).
- Detecting lexical semantic change (Bizzoni et al. 2019; Devlin et al. 2019; Giulianelli et al. 2020; Gong et al. 2020; Kutuzov et al. 2018; Peters et al. 2018; Sanh et al. 2019; Schlechtweg et al. 2020; Tahmasebi et al. 2019; Tsakalidis and Liakata 2020).
- (Diachronic) ontology learning from text (Asim et al. 2018; Bizzoni et al. 2019; Buitelaar et al. 2005; Gulla et al. 2010; He et al. 2014; Iyer et al. 2019; Rosin and Radinsky 2019; Wohlgenannt and Minic 2016).
- Documenting, "explainable AI" (Hyvönen 2020).

Tools

- Existing ontologies and linked data collections:
  Linguistic Linked Open Data Cloud, Linked Open Data Cloud
- Ontology learning tools and converters: CoW (Meroño-Peñuela et al. 2020); Fintan (Fäth et al. 2020); LODifier (Augenstein et al. 2012); LLODifier (Chiarcos et al. 2017, Cimiano et al. 2020); OntoGain (Drymonas et al. 2010); Text2Onto (Cimiano and Volker 2005).
- Semantic Web formalisms: RDF, OntoLex-Lemon, OWL-Time.
- SemEval 2020 task Unsupervised lexical semantic change detection (Schlechtweg et al. 2020).

Requirements

- Modelling semantic change via LLOD [support from T1.1]
- Generating and publishing multilingual parallel LLOD ontologies to trace the evolution of concepts [T1.2, T2.1, T2.5]
- Applying NLP methods (e.g. diachronic word embeddings) to detect and represent semantic change [T3.2]
- Linking parallel LLOD ontologies across different dimensions such as time, language and domain to facilitate multilingual and diachronic analysis of multifaceted concepts in the Humanities [T1.3, T3.4]
- Providing examples of applications, combining LLOD and diachronic analysis, that may be used in teaching linguistic data science [T3.5]

Support from other WGs and tasks (related to the possible connection points mentioned above) may also take the form of:

- shared expertise within Nexus (survey results, publications, state-of-the-art and WG/task reports, training schools, etc);
- direct involvement with the UC activities (group meetings and/or discussion groups on specific topics, paper proposals, presentations at conferences and workshops, experiments with various NLP and semantic Web technologies, models, languages and concepts, publication of LLOD ontologies, conception of methodological and/or pedagogical guidelines derived from the use case, Nexus joint reports or events, etc)

Languages

Ancient Greek, Hebrew, French, Latin, Old Lithuanian; Bulgarian, English, Polish, Romanian, Slovene (to be confirmed); other (to be defined).

Semantic fields, concepts

Socio-cultural transformation domain: geo-political and socio-cultural entities (including Europe, West, East, etc), education, sciences, technology and innovations, social and societal processes (migration, urbanisation, modernisation, globalisation), state and citizenship, beliefs, values and attitudes (e.g. religion, democracy, political participation), economy, health and well-being, everyday life, family and social relations, time and collective memory, work and leisure, customs and traditions, literature and philosophy.

Strategy

- The aim of the use case is to identify a set of rich and multifaceted concepts and semantic fields that are potentially interesting for comparative, multilingual and diachronic analysis (e.g. the domain of socio-cultural transformation, including *Europe* and related notions, *Western*, *Eastern*, *Orient*, *Occident*, etc), and to devise a methodology for tracing their evolution over time by means of NLP and LLOD technologies.
- The strategy will imply the use of resources in corpus linguistics, word embeddings-based approaches and Semantic Web formalisms during three main phases: (1) identify the concepts, languages, time span and datasets to be studied; (2) define and test the methodology for detecting semantic change (e.g. diachronic word embeddings) for the selected concepts and datasets; (3) generate a sample of multilingual parallel ontologies representing these changes and publish them as LLOD.
- The outcome will consist of a sample of multilingual parallel ontologies tracing the evolution of concepts and a set of guidelines describing the methodological approach applied in the use case.

Tasks

Task	Description
T0	Define use case and participation.
T1	Explore annotated diachronic corpora via specialised search engines and other relevant resources and define the set of concepts and languages to be analysed. Identify potential datasets to be used in the use case.
T2	Draw the SOTA in LLOD and NLP data/tools/methods for detecting and representing semantic change, with main application in the Humanities research. Define the general methodology of the use case, and the model for tracing historical change and the intended type(s) of semantic shifts, e.g. core (context-unspecific)/margin (context-specific) features, linguistic/cultural drifts.
T3	Select the datasets, periods and time span granularity (years, decades, centuries) and prepare the data to be used in change detection. This can include pre-processing (conversion from one format to another, cleaning, grouping by time period, etc) and preliminary exploration of the datasets with corpus linguistics tools (e.g. concordances, co-occurrences, specificities by time intervals), syntactic parsing, named entity recognition (NER) and semantic search engines.
T4	Study and choose the methods and tools for detecting semantic change and apply them to the selected data samples.
T5	Analyse T4 results and explore possibilities for semi-automatically generating ontological relations. Define the representation models and publish as LLOD the multilingual, parallel ontologies tracing the evolution of the target concepts.
T6	Document the whole process and produce a set of guidelines to describe the methodology derived from the use case.

Duration. 3 years, 4 months + 8 months (preparation).

Workflow

Task / Month	M6	M12	M18	M24	M30	M36	M42	M48
T0. Define use case and participation								
T1. Select concepts, languages								
T2. SOTA. Model for concept change								
T3. Prepare datasets and time slices								
T4. Study and apply change detection								
T5. Study and build LLOD ontologies								
T6. Document tasks, create guidelines								

Colour codes: - completed; - in progress; - not started

Methodology

The methodology involves a comparative, multilingual and interdisciplinary approach making use of various resources in areas such as corpus linguistics, word embedding and Semantic Web, as well as a selection of textual datasets in different languages and domains of the Humanities.

Deliverables

D0. Use case description (M8).

D1. Report on selected concepts, languages, models; model for representing concept change (M18).

D2. Selected datasets to be processed; report (M24).

D3. Change detection results; report (M36).

D4. LLOD published ontologies (M42).

D5. Final report and set of guidelines (M48).

Milestones

MS1. Theoretical framework of the use case (M18).

MS2. Selected datasets to be processed (M24).

MS3. Change detection results (M36).

MS4. LLOD published ontologies (M42).

MS5. Methodological guidelines (M48).

Collaboration and Exchange

- UC coordination and WG4 communication channels
- WG4 UCs and Tasks
- Nexus WGs

WG1 Linked data-based language resources (Task 1.1: LLOD modelling; Task 1.2: Creation and evolution of LLOD resources in a distributed and collaborative setting; Task 1.3: Cross-lingual data interlinking, access and retrieval in the LLOD);

WG2 Linked data-aware NLP services (Task 2.1: LLOD in knowledge extraction; Task 2.5: LLOD in terminology and knowledge management);

WG3 Support for linguistic data science (Task 3.2: Deep learning and neural approaches for linguistic data; Task 3.4: Multidimensional linguistic data; Task 3.5: Education in Linguistic Data Science).

- Other.
  - Possible participation in the ADHO SIG-LOD
  - Submission to CHANSE, Call for transnational research projects: *Transformations: Social and Cultural Dynamics in the Digital Age* (under evaluation)

Dissemination

- Reports
- Meetings, Workshops
- Conferences: DH, LDK, LREC, COLING, ISWC, SEMANTiCS, Semantic Web conferences and journals.
- Submitted papers:

  - *Semantic Web journal, Special Issue on Latest Advancements in Linguistic Linked Data:* LL(O)D and NLP Perspectives on Semantic Change for Humanities Research (major revision).

  - *LDK 2021 – 3rd Conference on Language, Data and Knowledge, 1-3 September in Zaragoza, Spain:* HISTORIAE, HIstory of Socio-culTural transfORmatIon as linguistIc dAta sciEnce. A Humanities Use Case (accepted).

UC 4.2.2. Use Case in Social Sciences

Overview. Survey data provide a valuable source of information and research for different scientific disciplines, such as social sciences, philosophy, anthropology, political sciences and history. They are also of interest for practitioners such as policy makers, politicians, government bodies, educators, journalists, as well as all other stakeholders with occupations related to people and society. Therefore, social data archives allowing open access to survey data are a crucial instrument for facilitating the use of these data for different purposes. The constitution of social data archives has to go together with language tools, allowing to find the necessary datasets, or to prepare them for research by third parties, and finally to make links between the data inside the different datasets in a given social data archive. Such tools consist of data anonymization tools, semantic search, semantic data integration and relations detection. Furthermore, data from social data archives can be linked with evidence about particular language phenomena and public attitudes that are found in the social media, such as language of aggression, or political preferences' influence, to provide a broader picture about the clusters of social attitudes. This use case is about building a toolset of language processing tools that enable the usage of survey data archives, organized according to linked data principles and providing generalizations about social attitudes clusters based on social media analysis and linking.

UC4.2.2. coordinator

Mariana Damova holds a PhD from the University of Stuttgart and is CEO of Mozaika, a company providing research and solutions in the field of data science. Her background is in Natural Language Processing, Semantic Web technologies and AI, with a strong academic and industrial record in North America and Europe.

The State-of-the-Art

Survey open questions provide free text answers that allow us to understand the person's opinion or attitude towards certain topics. These free text answers are valuable because they help profile the people taking the survey and grasp the reasons for the expressed opinions. Free text answers of surveys have many imperfections. They are usually messy, with grammar errors, spelling mistakes, and colloquialisms, and they come in high volumes. That is why natural language processing techniques are to be employed to make the analysis of the free answers easier. The most common points of interest in free answer-analysis are the detection of its topic, followed by opinion mining and sentiment analysis. To do this, approaches with different levels of complexity have been developed. Here are several examples:

- **Word Clouds**. Using the "bag of words" concept or building a specific dictionary of words, concepts and stems

- **Network Analysis**. Creating lists of topics of interest and then representing their relationships based on their occurrences in the texts, by visualising them as a graph of words (Figure 1).

Figure 1

- **Word Frequencies**. Counting the occurrences of the different words and phrases to produce word frequencies maps, clusters
- **TF-IDF (term frequency-inverse document frequency) matrix**. Allowing more complex analyses by downweighing terms that appear in all free text answers, while upweighting rare words that may be of interest
- **Clustering**. Using machine learning algorithms, such as K-means algorithm, to group the free text answers into distinct clusters
- **Latent Dirichlet Algorithm (LDA)**. Generating topics directly from the free text answers, using algorithms like the latent dirichlet algorithm
- **Sentiment analysis**. Identifying the polarity of the sentiment in the free text answer towards a given topic – positive or negative, or in more sophisticated cases - sentiment nuances, such as aspect-based sentiments, or scales of sentiments, or emotions with different approaches from sentiment lexicon based on machine learning (Abirami et al. 2016; OpenCV 2017; Sayad 2010) and ontology-based ones (Polpinij 2008; Gomez-Perez et al. 2002) that detect sentiments at whole text level, at sentence level or at attribute level
- **Opinion mining**. Understanding the drivers behind why people feel the way they do about a certain topic, subjectivity or bias analysis, helping to expose critical areas of strengths and weaknesses of the topic and tapping into the universe of unstructured opinion data to make better policy- and business-critical decisions, being regular opinions, expressing an attitude towards a subject matter or an attribute or comparative opinions, comparing two subject matters or attributes with machine learning, lexicon-based, corpus-based, dictionary-based approaches (Othman et al. 2014)

In more general terms, natural language processing for social sciences deals with creating methods to detect and identify language features indicating social attitudes, such as group decision making, viral moments during certain events, respectfulness, sentiment patterns, perceptions in the public sphere, moral classification, etc. All these topics have been addressed at the 2019 ACL Workshop.

Linked Open Data Technologies in the social sciences have been adopted to primarily link survey datasets, enabling the exploration of topics like "Are there non-elite electorate and if yes, where do they live?", using vocabularies about occupations (HISCO) (van Leeuwen et al. 2002), and about religions (LICR) to enrich the linked data datasets. Another application of Linked Open Data Technologies in the social sciences is enriching statistical analysis with linked data (Zapilko et al. 2011). Finally, Linked Open Data Technologies are used to describe the catalogues of social data repositories, like in the CESSDA.eu catalogue. However, semantic annotation techniques and use of Linked Open Data Technologies to interpret surveys or free text answers to open questions have not been adopted so far.

Resources

Survey data are available in open access repositories, e.g.:

1. CESSDA.eu - an umbrella organization where surveys from all over Europe are collected.

2. FORSCENTER.ch – the Swiss centre of expertise in social sciences

3. Local ecosystems' survey data, and survey data catalogues

Social media corpora about different topics provided by the participants:

1. Speech of aggression

2. Political preferences towards politicians

3. Study of social inequalities in transition from school to the job market

Language resources. Different vocabularies have to be established:

1. Discourse markers

2. Attitude vocabularies

3. Opinion, sentiment and topics vocabularies

Datasets from the Linked Open Data cloud will be reused. Ontologies will be developed and LLOD resources will be adopted.

Methods

As the goal of the use case is to collect methods for appropriate processing of free text answers to open questions in surveys about social inequalities, as well as regional difference in the transition from school to work force, opinions about politicians, and aggressive language from social media, we will explore different state of the art approaches, listed in the SOTA section and evaluate them in order to provide specification of the proper application area of the given method. The evaluation of the methods will depend on the selected corpora/datasets and their curation. We will devise workflows and guidelines for the adoption of language processing approaches depending on the datasets to be targeted. Moreover, we will elaborate workflows for datasets curation, including data anonymisation techniques (Kleinberg et al. 2017; Mosallanezhad et al. 2019), and user profiling. Furthemorer, we will establish guidelines for the creation of LLOD vocabularies for discourse markers, aggressive expressions, favourable or unfavourable attitude expressions, topics

descriptions, and apply for funding to create and publish such LLOD vocabularies. In addition, we will analyse the links between survey analysis and social media analysis. In the analysis of survey datasets, we will explore the impact of dialogue modelling (Su et al. 2019) and question answering techniques (Soares and Parreiras 2020) for better interpretation of the free text answers to open questions from the surveys and maybe full surveys.

Tools

Apart from the approaches listed in the SOTA section and in the methodology section, we singled out freely available language processing tools for social sciences that we will evaluate. For example, the NLP tools for social sciences website (Crossley et al. 2014) puts together freely available tools that measure parameters related to lexical sophistication, text cohesion, syntactic complexity, lexical diversity, grammar/mechanics and sentiment analysis.

SiNLP. The Simple Natural Language Processing Tool allows users to analyse texts using their own custom dictionaries. In addition to analysing custom dictionaries, SiNLP also provides the name of each text processed, the number of words, number of types, Type-Toke-Ratio (TTR), Letters per word, number paragraphs, number of sentences, and number of words per sentence for each text. Included with SiNLP is a starter custom list dictionary that includes determiners, demonstratives, all pronouns, first person pronouns, second person pronouns, third person pronouns, conjuncts, connectives, negations, and future.

Text analysis uses NLP to automate the process of classifying and extracting data from texts, such as survey responses, product reviews, tweets, emails, and more. In other words, it automatically structures your data and allows you to get insights about your business. The University of Oxford offers a course in NLP for social sciences, treating tools for large-scale analysis of linguistic data such as document collections, transcripts, and blogs, based on statistical principles such as Naïve Bag of Words, but also on effects of social and pragmatic context, clustering, classifying based on words sequences to characterize the topics of different documents, as well as the socio-indexical traits of the speakers or the authors, ultimately to analyse the spread of memes and opinions through repeated interactions in linguistic communities.

MonkeyLearn has a number of pre-trained models that can help one analyse one's survey results right away. For example, the sentiment analysis model will help see if one's customers' responses are *Negative*, *Positive*, or *Neutral*, while the aspect classifier identifies the theme or topic those customers mention.

SPSS Analytics Partner IBM SPSS Text Analytics for Surveys uses powerful NLP technologies specifically designed for survey text. It leads the way in unlocking open-ended responses for better insight and statistical analysis. IBM SPSS Text Analytics for Surveys categorizes responses and integrates results with other survey data for better insight and statistical analysis, automating the categorization process to eliminate the time and expense of manual coding, and using linguistics-based technologies to reduce the ambiguities of human language, helping one uncover patterns in the attitudes, beliefs and opinions of others.

Perceptix uses NLP For Open-Ended Survey Questions Analysis to detect sentiment and topics in the free text answers. Sentiment analysis of positive, negative, and neutral responses is used to flag areas where more information is needed; a high negative score serves as a cue to drill deeper to determine the cause of discontent. Recurring themes or topics are also a flag to signal what is on the minds of most surveyed people and may need more study.

The **ELG** platform provides a number of language processing technologies based on semantics and language resources that offer a rich library of instruments for survey analysis to evaluate.

Requirements

WG1 support in the creation of vocabularies of discourse markers, attitude vocabularies, opinion, sentiment and topics vocabularies and LLOD models for them

WG2 support in survey data collection from CESSDA.eu and other local sources, and multilingual parallel corpora constitution

WG3 support in stakeholder requirements collection and multilingual corpora constitution in English, Hebrew, Bulgarian, Latvian, Polish, German and other languages in the LLOD cloud

Languages

English, Hebrew, Bulgarian, Latvian, Polish and others

Roadmap

Figure 2 shows the roadmap for the execution of the WG4 Social Sciences use case. It is devised including five consequent and interdependent steps:

- Collection of stakeholders and requirements
- Selection and constitution of Survey corpora
- Selection and evaluation of NLP tools and resources
- Specification of LLOD and LOD representation guidelines
- Building prototypes and research project proposals

Figure 2

Strategy

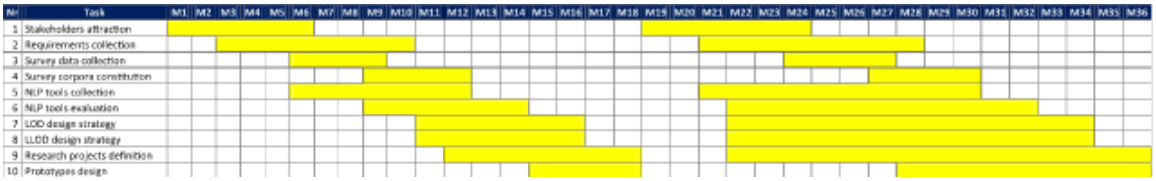
- research collaboration within the interested researchers in the Social Sciences use case,
- within the NexusLinguarum WGs, and external stakeholders and data providers
- identification and re-use of suitable methodologies, approaches and tools
- implementing best practices of LLOD and LOD development

Tasks

No.	Task	Description
1	Stakeholders attraction	Identification, contact, awareness raising and attraction of stakeholders
2	Requirements collection	Interviewing stakeholders and definition of users, specification of requirements
3	Survey data collection	Collection of surveys corresponding to the topics of interest
4	Survey corpora constitution	Analysis of the collected surveys and constitution of corpora in easy for processing format
5	NLP tools collection	Selection of NLP tools corresponding to the topics of interest and to the requirements
6	NLP tools evaluation	Evaluation of the selected NLP tools
7	LOD design strategy	Analysis and definition of the adoption of LOD for Survey processing based on the defined requirements
8	LLOD design strategy	Analysis and definition of the adoption of LLOD for Survey processing based on the defined requirements
9	Research projects definition	Definition of research topics, formation of project consortia, submission of research proposals
10	Prototype design	Description of guidelines for developing resources, tools or solutions for surveys processing with LLOD and LOD methods

Duration. From July 2020 – June 2023 (36 months)

Workflow



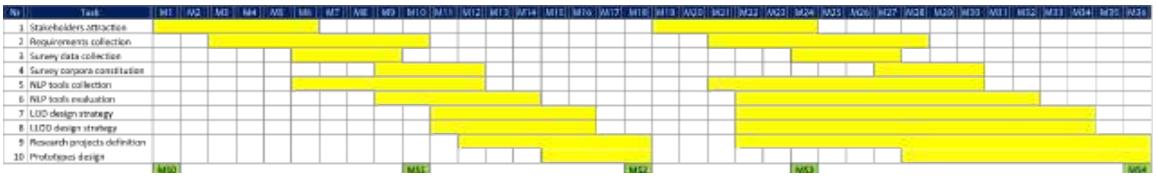
Deliverables

D1. Initial Use case design - M18

D2. Intermediary Use case design - M24

D3. Final Use case design - M36

Milestones



Collaboration and Exchange

- UC coordination and WG4 communication channels - Slack
- WG4 UC 4.1.1.
- Nexus WG1, WG2
- STSMs (to be defined)
- CESSDA, FORSCENTER, University of Gothenburg, Bulgarian Academy of Sciences and others
- Consortia for H2020 and other bilateral funding for R&D opportunities
- Dedicated Slack channel for the WG4 Use case in Social Sciences
- Bi-weekly meetings

Dissemination

- Reports
- Meetings, Workshops, Conferences
- Publications

Task 4.3. Use Cases in Technology

Task 4.3. builds upon the recent advancements in the areas of multilingual technologies, machine translation, automatic term extraction methods, text analytics and sentiment analysis models with the aim to reuse existing open-source components and test them in different Information and Communication Technologies and business scenarios. General subtasks within this task include: state-of-the-art analysis; requirements elicitation and use case definition; compilation of corpora, term extraction and semantic linking, document classification; and evaluation of NLP tools in different scenarios. During the first year, two specific Use Cases have been selected: Cybersecurity and FinTech. The emphasis of the Cybersecurity use case (UC4.3.1.) is on terminology extraction, with the goal of compiling a bilingual/multilingual term base of cybersecurity terms and their metadata in at least two languages. The emphasis of the FinTech use case (UC4.3.2.) is on sentiment analysis (SA), with the goal of developing domain-specific SA models that can provide an efficient method for extracting actionable signals from the news. Activities in both scenarios are coupled with running national and commercial projects and thus will have impact on researchers and industrial users of language technologies.

UC 4.3.1. Use Case in Cybersecurity

UC 4.3.2. Use Case in Fintech

T4.3. leader, linguistics

Daniela Gîfu is a Scientific Researcher at the Faculty of Computer Science of Alexandru Ioan Cuza University of Iasi. Her main interests include corpora, sentiment analysis, semantic web, machine and deep learning, AI and system development. She has been involved in research projects and publications.

T4.3. leader, computational

Valentina Janev is a Senior Researcher at the Institute ˝Mihajlo Pupin˝, University of Belgrade. Her research interests include business information systems, AI, semantic web technologies, knowledge-based approaches to NLP, etc. She has authored and edited books and numerous papers, and is involved in the coordination of H2020 projects.

UC 4.3.1. Use Case in Cybersecurity

Overview. The aim of the use case is to develop a methodology for compilation of a termbase for the cybersecurity (CS) domain using deep learning systems and LLOD principles. The datasets encompass both parallel and comparable corpora, providing the possibility to extract terms not only from original texts and their translations, but also from comparable original texts of the same domain in several languages. This methodology is believed to be highly suitable for under-resourced languages, as it expands the amount and variety of data sources which can be used for term extraction. State-of-the-art neural networks will be developed and applied for automatic extraction of terminological data and metadata necessary for termbase compilation.

UC4.3.1. coordinator

Sigita Rackevičienė is a professor at the Institute of Humanities of Mykolas Romeris University and senior researcher in the Lithuanian national scientific project Bilingual Automatic Terminology Extraction. Her scientific interests encompass multilingual terminology and terminography.

Resources.

- EURLex (for parallel corpus),
- national and international legislation, public documents of national and international cybersecurity institutions, academic literature, specialised and mass media (for comparable corpus)

Methods.

- parallel and comparable corpora building methodology;
- manual terminology annotation and development of gold standard corpora;
- automatic bilingual term extraction and alignment using deep learning systems and gold standard datasets as training data;
- automatic extraction of knowledge-rich contexts using deep learning systems;
- development of an interlinked termbase using LLOD.

Tools

Manual monolingual and bilingual terminology annotation software, neural networks for automatic data extraction.

Requirements

- Compilation of parallel and comparable bilingual cybersecurity corpora and termbase
- Development of small-scale gold standard bilingual corpora with manually annotated terms for training and assessment of neural network systems
- Development of neural network systems for bilingual automatic extraction of terminological data and metadata

WG1 support in applying LLOD technologies for interlinking the compiled termbase with other resources with regards to the models/best practices surveyed or under development in the context of T1.1 and T1.2, plus the exploration of techniques for interlinking under analysis in T1.3.

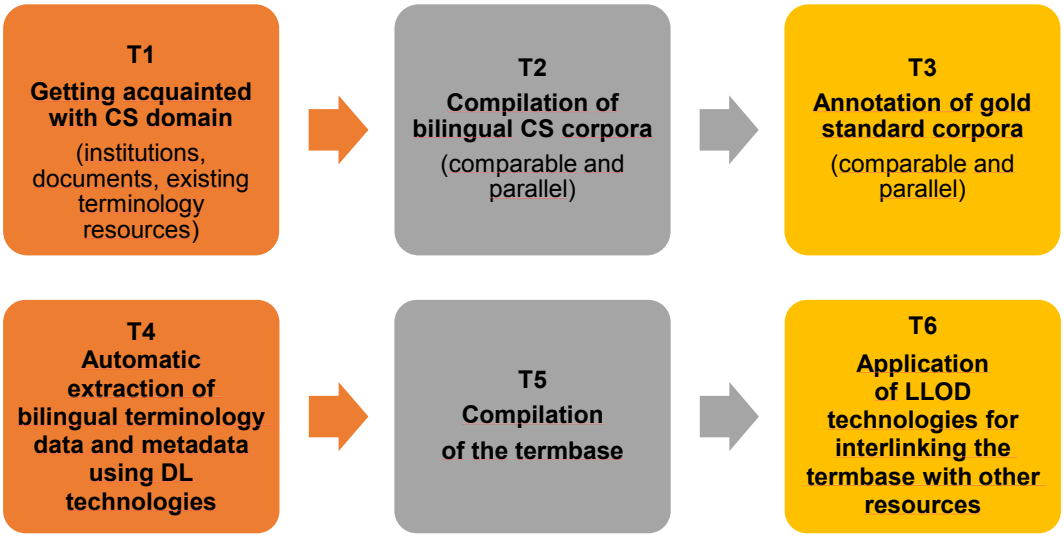
Languages

English and Lithuanian.

Strategy and Roadmap

The main objectives encompass the development of methods to allow regular update of termbases by automatically extracting terminology and knowledge-rich contexts from relevant texts, as well as integration of the compiled terminological data into the global LLOD ecosystem.

Tasks



T1: Getting acquainted with CS domain: searching and getting acquainted with existing English cybersecurity termbases, glossaries, ontologies; systematisation of the collected information.

T2: Compilation of bilingual CS corpora - building a knowledge store:

T2.1. Examination of international CS documents which are translated into other languages, their collection and compilation of a parallel corpus (the EU legislative documents in EURLex database; international conventions, etc.)

T2.2. Examination of national CS documents, their collection and compilation of a comparable corpus (national legal acts and administrative documents; academic texts; specialised and mass media texts, etc.).

T3. Annotation of gold standard corpora for training and assessment of machine learning and neural network systems

T4: Automatic extraction of bilingual terminology data and metadata using deep learning technologies:

T4.1. Iterative testing of the automatic term extraction methods by comparing their results with the gold standards;

T4.2. Selection of the most effective methods and automatic extraction of term candidates from parallel and comparable corpora, their automatic alignment;

T4.3. Selection of the dominant CS terms based on frequency/dispersion analysis and expert approval.

T4.4. Development of automatic methods for extraction of knowledge-rich contexts; their extraction for the selected dominant CS terms.

T5: Compilation of the termbase:

T5.1. Formulating final definitions of the terms using the extracted knowledge-rich contexts;

T5.2. Collecting other metadata about the selected terms: usage examples; conceptual relations with other terms, statistical data on term frequency and dispersion, etc.

T5.3. Uploading the collected data to a termbase.

T6: Application of LLOD technologies for interlinking the termbase with other resources.

Workflow

	2020		2021		2022		2023	
Task / Months	**M6**	**M12**	**M18**	**M24**	**M30**	**M36**	**M42**	**M46**
T0. Defining use case and participation								
T1. Getting acquainted with CS domain								
T2. Compilation of bilingual CS corpora								
T3. Annotation of gold standard corpora								
T4. Automatic extraction of bilingual terminology data and metadata								
T5. Compilation of the termbase								
T6. Application of LLOD technologies for interlinking the termbase with other resources								

Colour codes: - completed; - in progress; - not started

Deliverables

D1. Parallel and comparable corpora of the CS domain, made available to the public in the CLARIN repository.

D2. Termbase of CS terms. The base will be publicly available on the internet.

D3. Publications on the results of the use case.

Collaboration and Exchange

- UC coordination and WG4 communication channels
- cooperation with other UCs in WG4 and with other Nexus WGs

Dissemination

Reports, meetings, workshops, conferences, publications

Acknowledgement

The use case is based on the project “Bilingual automatic terminology extraction” funded by the Research Council of Lithuania (LMTLT, agreement No. P-MIP-20-282).

UC 4.3.2. Use Case in Fintech

UC4.3.2. coordinator

Dimitar Trajanov heads the Department of Information Systems and Network Technologies at the Faculty of Computer Science and Engineering, ss. Cyril and Methodius University, Skopje. He has been involved in more than 60 projects and is the author of 160 journal and conference papers.

Overview. Financial systems are among the most dynamic and innovative systems in the world, used for financial services, markets, banks, corporations, central banks, investors, traders, brokers, dealers are diverse participants in the financial system who influence its dynamics.

Among other disciplinary approaches to study financial markets, computational linguistics has become increasingly powerful due to the availability of large text datasets pertaining to the determinants of financial market performance and individual companies' prospects. The development of increasingly powerful methodologies for text analytics has contributed to improvement in NLP techniques.

Sentiment Analysis (SA) is one of the most important applications of NLP in finance, allowing prompt extraction of positive or negative sentiments from the news as support for decision making by traders, portfolio managers and investors. SA models can provide an efficient method for extracting actionable signals from the news. General SA models are ineffective when applied to specific domains such as finance, so the development of domain-specific models is needed. In this use case, the overview of the models and application of Sentiment Analysis in the finance domain will be presented.

The State-of-the-Art

The financial domain is characterised by unique vocabulary which calls for domain-specific sentiment analysis. The sentiments expressed in news and tweets influence stock prices and brand reputation, hence, constant measurement and tracking these sentiments is becoming one of the most important activities for investors.

Given that the financial sector uses its own jargon, it is not suitable to apply generic sentiment analysis in finance because many of the words differ from their general meaning. For example, "liability" is generally a negative word, but in the financial domain, it has a neutral meaning. The term "share" usually has a positive meaning, but in the financial domain, a share represents a financial asset or a stock, which is a neutral word. Furthermore, "bull" is neutral in general, but in finance, it is strictly positive, while "bear" is neutral in general, but negative in finance. These examples emphasise the need for the development of dedicated models, which will extract sentiments from financial texts.

Resources

- Dataset: The Financial Phrase-Bank consists of 4845 English sentences selected randomly from financial news found on the LexisNexis database

- Dataset: SemEval-2017 task "Fine-Grained Sentiment Analysis on Financial Microblogs and News". Financial News Statements and Headlines dataset consists of 2510 news headlines, gathered from different publicly available sources such as Yahoo Finance
- Dataset: bank-additional-full.csv consists of 41188 data points with 20 independent variables, 10 of which are numeric features and the remaining 10 are categorical features, ordered by date (from May 2008 to November 2010), very close to the data analyzed in Moro et al. 2014

Methods

- Lexicon-based approaches for sentiment analysis in finance
- Statistical feature extraction from texts without external knowledge
- Word representation methods
- Sentence encoders
- NLP models based on the transformer neural network architecture

Tools

- Classification models: SVM, Neural Network, XGBoost
- Fine tuning of pretrained transformer models

Requirements

- Identify the methods and algorithms that are used for Sentiment Analysis in finance
- Evaluate the different approaches to find the best one(s) for specific tasks in finance
- Identify potential applications of sentiment analysis models in different finance-related activities

Languages

English (others will be added throughout the Action)

Strategy

- The aim of the use case is to identify the methods and algorithms that can be used for Sentiment analysis in Finance
- Evaluate the different approaches in order to find the best one for specific tasks in finance
- Identify potential applications of sentiment analysis models in different finance-related activities

Tasks

T0. Define use case and participation

T1. State-of-the-art

T2. Evaluate the different approaches

T3. Find potential applications

T4. Expand the model for other languages

Duration: 4 years

Workflow

	M6	M12	M18	M24	M30	M36	M42	M48
T0. Define use case and participation								
T1. State-of-the-art								
T2. Evaluate the different approaches								
T3. Find potential applications								
T4. Expand for other languages								

Deliverables

D0. Use case description

D1. Report on selected concepts, languages, models

D2. Evaluation of the results

D3. Application in other languages

D5. Final report and set of guidelines

Milestones

MS1. Survey of the current approaches

MS2. Evaluation of the models

MS3. Application of the created models

Collaboration and Exchange

- UC coordination and WG4 communication channels
- WG4 UCs and Tasks
- STSMs

Dissemination

- Meetings, Workshops
- Nexus activities
- Conferences, Publications

Journals (selection)

- *Business & Information Systems Engineering* (impact factor = 3.6)
- *Business Process Management Journal* (rank B on CORE Platform; impact factor = 1.46)
- *Business Intelligence Journal* (rank C_CORE Platform)

Task 4.4. Use Cases in Life Sciences

Overview. The area of Life Sciences is broad and heterogeneous. For that reason, the task T4.4. is constrained to a general overview and focused investigation of three important subtopics: *Public Health, Ecology,* and *Pharmacy*. Our investigation targets in particular disease prevention and quality of life.

The task aims to cover the above-mentioned life science topics within news media and social media in a cross-lingual setting. The main information sought concerns the COVID-19 pandemic situation. However, we will add other sources of information, including scientific literature on life sciences and its relation with linked data.

Until March 2021, T4.4. has evolved in a general manner. Then, it was decided to divide it into a Use Case in Public Health (UC4.4.1.) to handle the crux of this task, as described in this section, and to form another Use Case in Pharmacy (UC 4.4.2.). In addition, it is planned to form a new Use Case on Ecology (UC4.4.3.).

T4.4. leader, linguistics
(until June 2021)

Petya Osenova is senior researcher in Language Technologies at the Department of AI and LT of the Institute of Information and Communication Technologies, Bulgarian Academy of Sciences. Her scientific interests are in the fields of formal grammars and computational linguistics, language resources, language modeling and machine translation.

T4.4. leader, linguistics
(from June 2021)

Ana Ostroški Anić is a Research Associate at the Institute of Croatian Language and Linguistics. Her background is in linguistics and terminology management, with research interests including aviation terminology, medical terminology, specialized knowledge processing, frame semantics and metaphor research.

T4.4. leader, computational

Marko Robnik-Šikonja is Professor of Computer Science and Informatics at University of Ljubljana, Faculty of Computer and Information Science. His research interests span machine learning, data mining, NLP, network analytics, and application of data science techniques. He is (co-)author of over 150 scientific publications that were cited more than 5,500 times.

Resources

We rely on several types of resources: available ontologies, corpora and lexical databases (such as terminological dictionaries). Other resources specifically related to COVID-19 data include:

- EU Open Data Portal
- Novel Coronavirus (COVID-19) Cases Data
- 5 Datasets About COVID-19
- BioPortal: A dataset of linked biomedical ontologies
- WikiData on COVID-19
- International Statistical Classification of Diseases and Related Health Problems

Methods

When data is identified and gathered, as well as the related ontologies and lexicons, the following methods are applied: Machine Learning, Information Extraction, and NLP. The linguistic pipelines such as Stanza cover most European languages and provide the baseline text processing, such as tokenization, lemmatization, part-of-speech (POS)-tagging, named entity recognition, and to a lesser degree dependency parsing.

Technologies and Approaches

Open Data, Embeddings, Knowledge Graphs, Deep neural networks.

We rely on the pre-trained word embeddings for mono and multilingual settings, as well as on the existing linked data (domain ontologies, Wikipedia, specialized thesauri). For prediction models we use monolingual and multilingual variants of large pretrained language models, based on the transformer neural networks, such as BERT and RoBERTa models.

Requirements

WG1 support in Knowledge Resources, including specialized corpora in Life Sciences or related data that contains such information, terminological dictionaries, lexical databases, ontologies (preferably LLOD)

WG2 support in Technology (Tools) for information extraction and explainable analytics, such as linguistic/stochastic pipelines that can handle knowledge rich data, pre-trained embeddings for low-resourced languages, etc.

WG3 support in preparing data sets in Public Health, Pharmacy and Ecology (Data Management)

Languages

We cover English and will feature news and social media in cross-lingual settings, focusing on less-resourced languages, e.g., Slovene, Bulgarian, Portuguese, Macedonian, or Croatian.

Strategy

We began with an informative survey of the SOTA in the selected topics, covering specific resources, methods, technologies and approaches. Based on that, we will identify opportunities, collect datasets and perform initial analyses involving knowledge extraction and information retrieval.

Tasks

T1. State-of-the art overview: general trend in life sciences, in public health and Ecology and in pharmacy

T2. Identification of the related resources and tools

T3. Description of their status (advantages, problems, etc)

T4. Preparation of datasets

T5. Analytics: information retrieval, knowledge extraction and explanation of models

Duration. From July 2020 to October 2023

Workflow

Task/Month	M1	M6	M7	M12	M13	M19	M21	M22	M23	M28
Task 1										
Task 2										
Task 3										
Task 4										
Task 5										

Deliverables

D1. Deliverable on SOTA (M8)

D2. Deliverable on identification and description of related resources and tools (M19)

D3. Deliverable on datasets and analytics (M28)

Milestones

MS1. State-of-the-art (M7)

MS2. Identification of language resources (M12)

MS3. Description of available LRE (M19)

MS4. Datasets (M22)

MS5. Probes on Analytics Approaches (M28)

Collaboration and Exchange

- WG4 – expertise in working with social media and regarding technology
- Nexus WGs – interaction with all other WGs
- STSMs

Dissemination

- Reports – all planned deliverables will serve also as reports
- Meetings, workshops, conferences

Interaction with the other Working Groups

Since one of WG4's core goals is putting into practice the various resources and techniques studied and developed in the other Nexus WGs, this has implied building up different forms of interaction between WG4 and the other WGs in general, as well as between the WG4 UCs and other WG Tasks in particular.

The following table offers an overview of these interactions established by Q2 of 2021.

	UC4.1.1.	UC4.1.2.	UC4.2.1.	UC4.2.2.	UC4.3.1.	UC4.3.2.
T1.1.	■		■	■	■	
T1.2.			■		■	
T1.3.			■		■	
T1.4.						
T1.5.					■	
T2.1.			■		■	
T2.2.						
T2.3.						
T2.4.						
T2.5.			■		■	
T3.1.						
T3.2.			■	■	■	
T3.3.					■	
T3.4.	■	■	■			
T3.5.			■			

WG1 - LD-based LRs	
T1.1.	Modelling
T1.2.	Resources
T1.3.	Interlinking
T1.4.	Sources quality
T1.5.	Under-resourced languages

WG2 - LD-aware NLP services	
T2.1.	Knowledge Extraction
T2.2.	Machine Translation
T2.3.	Multilingual Question-Answering
T2.4.	WSD & Entity Linking
	Terminology & Knowledge
T2.5.	Management

WG3 - Support for LD science	
T3.1.	Big Data & linguistic information
T3.2.	Deep Learning & neural approaches
T3.3.	Linking structured multilingual data
T3.4.	Multidimensional linguistic data
T3.5.	Education in Linguistic Data Science

WG4 - Use cases and applications	
UC4.1.1.	Media and Social Media
UC4.1.2.	Language Acquisition
UC4.2.1.	Humanities
UC4.2.2.	Social Sciences
UC4.3.1.	Cybersecurity
UC4.3.2.	FinTech

Concluding remarks and next steps

This overview shows that the current WG4 structure is stable, with a variety of UCs across relevant domains (Media and Social Media, Language Acquisition, Humanities, Social Sciences, Cybersecurity, FinTech, Public Health, and Pharmacy). It is believed that the results from upcoming calls for new UCs will further strengthen this foundation. The fact that more than 80% of the Action members participate in this WG constitutes a good basis for ongoing and future work by broadening the number of analysed languages, with a special emphasis being placed on under-resourced languages.

The added value of bringing together members with various backgrounds, particularly linguistic and computational, provides intra-WG expertise which helps foster internal collaboration but also facilitates inter-WG exchanges. As described in the Action's Memorandum of Understanding, WG4 aims to provide application scenarios to test the tools, technologies, and methodologies developed across NexusLinguarum. At the moment, some challenges related to modelling and interlinking (WG1) are already being addressed within the UCs, as well as issues concerning multidimensional linguistic data and deep learning (WG3).

It is expected that as work continues unfolding in WG4, and collaboration across WGs and Tasks becomes more frequent, additional topics can be further explored and results disseminated via joint publications. In the current scenario of travel restrictions due to the COVID-19 pandemic, and in case this situation persists, it would be very important that virtual Short-Term Scientific Missions (STSMs) could be implemented to support these collaborative interactions, which are, ultimately, one of the axes underpinning COST Actions.

References

Abirami, M.A.M. and Gayathri, M.V. 2016. A survey on sentiment analysis methods and approach. 2016 Eighth Int. Conf. Adv. Comput. 72–76, 10.1109/IcoAC.2017.7951748

Agaian, S. and Kolm, P. 2017. Financial sentiment analysis using machine learning techniques. *International Journal of Investment Management and Financial Innovations*, 3: 1–9.

Asim, M.N. Wasim, M. Khan, M.U.G. Mahmood, W. Abbasi, H.M. 2018. *A Survey of Ontology Learning Techniques and Applications*. Database 2018 January 1, 2018. https://doi.org/10.1093/database/bay101.

Atzeni, M., Dridi, A. and Recupero, D. R. 2017. Fine-grained sentiment analysis of financial microblogs and news headlines. *Semantic Web Evaluation Challenge*. 124–128.

Augenstein, I. Padó, S. and Rudolph, S. 2012. LODifier: Generating Linked Data from Unstructured Text, volume 7295 of Lecture Notes in Computer Science, 210–224. Berlin Heidelberg: Springer. http://link.springer.com/10.1007/978-3-642-30284-8_21, doi:10.1007/978-3-642-30284-8_21.

Betti, A. and Van den Berg, H. 2014. Modelling the History of Ideas. *British Journal for the History of Philosophy* 2: 812–35. https://doi.org/10.1080/09608788.2014.949217.

Bizzoni, Y. Mosbach, M. Klakow, D. and Degaetano-Ortlieb, S. 2019. Some Steps towards the Generation of Diachronic WordNets. *Proceedings of the 22nd Nordic Conference on Computational Linguistics NoDaLiDa. 55–64 Turku, Finland, 30 September – 2 October, 2019*. Linköping University Electronic Press, 10.

Buitelaar, P. Cimiano, P. and Magnini, B. 2005. Ontology learning from text: An overview. In *Ontology Learning from Text: Methods, Evaluation and Applications*, 123: 3–12.

Chiarcos, C., Ionov, M., Rind-Pawlowski, M., Fäth, C.,Wichers Schreur, J., and Nevskaya, I. 2017. LLODify-ing Linguistic Glosses. In Language, Data, and Knowledge. LDK 2017. *Lecture Notes in Computer Science*, 10318: 89–103. Cham: Springer.

Cimiano, P., Chiarcos, C., McCrae, J.P. and Gracia J. 2020. *Linguistic Linked Data in Digital Humanities*. In Linguistic Linked Data. Cham: Springer. https://link.springer.com/chapter/10.1007/978-3-030-30225-2_13.

Cimiano, P. and Volker, J. 2005. Text2Onto. A Framework for Ontology Learning and Data-driven Change Discovery. Montoyo, A., Munoz, R, and Metais, E. (eds.). *Natural Language Processing and Information Systems: 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Alicante, Spain, June 15 – 17, 2005; Proceedings*. Lecture notes in computer science, 3513. Springer: 227-238.

Crossley, S.A. and McNamara, D.S. 2009. Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing* 18: 119-135.

Crossley, S.A., and McNamara, D.S. 2010. Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson and R. Catrambone, (eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 984-989. Austin, TX: Cognitive Science Society.

Crossley, S.A., Roscoe, R., Graesser, A. and McNamara, D.S. 2011. Predicting human scores of essay quality using computational indices of linguistic and textual features. In G. Biswas, S. Bull, J. Kay and A. Mitrovic, (eds.*), Proceedings of the 15th International Conference on Artificial Intelligence in Education* 438-440. Auckland, NZ: AIED.

Crossley, S.A., Allen, L.K., Kyle, K. and McNamara, D.S. 2014. Analyzing discourse processing using a simple natural language processing tool SiNLP. *Discourse Processes*, 515-6. 511-534, DOI: 10.1080/0163853X.2014.910723

Davidson, T., Warmsley, D., Macy, M.W. and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the 11th International Conference on Web and Social Media ICWSM) 2017, Montréal, Québec, Canada, May 15-18, 2017*, 512-515. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665

Devlin, J. Chang, M.-W. Lee, K. and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186. Association for Computational Linguistics. Doi:10.18653/v1/N19-1423.

Dodevska, L., Petreski, V., Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. and Trajanov, D. 2019. Predicting companies stock price direction by using sentiment analysis of news articles. *15th Annual International Conference on Computer Science and Education in Computer Science*.

Drymonas, E. Zervanou, K. and Petrakis, E.G.M. 2010. Unsupervised Ontology Acquisition from Plain Texts: The OntoGain System. *Volume 6177 of Lecture Notes in Computer Science*, 277–287. Berlin Heidelberg: Springer. URL: http://link.springer.com/10.1007/978-3-642-13881-2_29, doi:10.1007/978-3-642-13881-2_29.

Ehrmann, M. Romanello, M. Clematide, S. Ströbel, P. and Barman, R. 2020. Language resources for historical newspapers: The Impresso collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*. *European Language Resources Association ELRA*.

Fäth, C. Chiarcos, C. Ebbrecht, B. and Ionov, M. 2020. Fintan – flexible, integrated transformation and annotation engineering. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, 7212–7221. European Language Resources Association ELRA.

Fergadiotis, G., Wright, H. and Green, S. 2015. Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research* 583, 840-852. Doi: 10.1044/2015_JSLHR-L-14-0280

Fokkens, A., Braake, S.T., Maks, I., Ceolin, D. 2016. On the Semantics of Concept Drift: Towards Formal Definitions of Concept Drift and Semantic Change. Drift-a-LOD@EKAW.

Geeraerts, D. 2010. *Theories of lexical semantics*. Oxford University Press.

Gelumbeckaite, J. Sinkunas, M. and Zinkevicius, V. 2012. Old Lithuanian Reference Corpus Sliekkas) and Automated Grammatical Annotation. *Journal of Language Technology and Computational Linguistics*, 272: 83–96.

Ghiassi, M., Skinner, J. and D. Zimbra 2013. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, 40.16: 6266–6282.

Giulianelli, M. Del Tredici, M. and Fernández, R. 2020. Analysing Lexical Semantic Change with Contextualised Word Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3960–3973. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.365.

Gomez-Perez, A. and Corcho, O. 2002. Ontology languages for the Semantic Web. *IEEE Intelligent Systems*, 17.1, 54-60. doi: 10.1109/5254.988453.

Gong, H. Bhat, S. and Viswanath, P. 2020. Enriching Word Embeddings with Temporal and Spatial Information. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, 1–11. Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.conll-1.1.

Graesser, A. C., McNamara, D. S., Louwerse, M.M., and Cai, Z. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 36, 193–202.

Gulla, J.A. Solskinnsbakk, G. Myrseth, P. Haderlein, V. and Cerrato, O. 2010. Semantic drift in ontologies. In *WEBIST 2010, Proceedings of the 6th International Conference on Web Information Systems and Technologies*, 2.

He, S. Zou, X. Xiao, L. and Hu, J. 2014. Construction of Diachronic Ontologies from People's Daily of Fifty Years. *LREC 2014 Proceedings*.

Howard, J. and Ruder, S. 2018. Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

Hyvönen, E. 2020. Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-Analysis and Serendipitous Knowledge Discovery. *Semantic Web Journal*. http://semantic-web-journal.net/content/using-semantic-web-digital-humanities-shift-data-publishing-data-analysis-and-serendipitous.

Iyer, V., Mohan, M. Reddy, Y.R.B. Bhatia, M. 2019. A Survey on Ontology Enrichment from Text. *The sixteenth International Conference on Natural Language Processing ICON-2019. Hyderabad, India*.

Johnson, R. and Zhang, T. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1: 562–570.

Khan, A.F. 2018. Towards the Representation of Etymological Data on the Semantic Web. Information 9, no. 12 November 30, 2018): 304. https://doi.org/10.3390/info9120304.

Khan, A.F. 2020. Representing Temporal Information in Lexical Linked Data Resources. In M. Ionov, J.P. McCrae, C. Chiarcos, T. Declerck, J. Bosque-Gil and J. Gracia (eds.). *Proceedings of the 7th Workshop on Linked Data in Linguistics LDL-2020, LREC 2020 Workshop, Language Resources and Evaluation Conference, 11–16 May 2020*.

Kim, Y. 2014. Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882.

Kleinberg, B., Mozes, M., van der Toolen, Y., and Verschuere, B. 2018. *NETANOS - Named entity-based Text Anonymization for Open Science*. Retrieved from osf.io/973rj

Kuukkanen, J-M. 2008. Making Sense of Conceptual Change. *History and Theory* 47.3, 351–72.

Kutuzov, A. Øvrelid, L. Szymanski, T. Velldal, E. 2018. Diachronic Word Embeddings and Semantic Shifts: A Survey. In *Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, August 20, 2018*. 1384–1397.

Kyle, K., Crossley, S. and Berger, C. 2018. The tool for the automatic analysis of lexical sophistication TAALES): version 2.0. *Behavior Research Methods*, 50.3, 1030–1046.

Li. N., Liang, X., Li, X., Wang, C. and Wu, D.D. 2009. Network environment and financial risk using machine learning and sentiment analysis. *Human and Ecological Risk Assessment*, 15.2, 227–252.

Liebeskind C. and Liebeskind, S. 2020. Deep Learning for Period Classification of Historical Hebrew Texts. *Journal of Data Mining and Digital Humanities*.

Loughran, T. and McDonald, B. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66.1, 35–65.

Luminoso. *Employee Feedback and Artificial Intelligence: A guide to using AI to understand employee engagement*. Accessed November 2018.

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

MacWhinney, B. 2020. *The CHILDES Project: Tools for Analyzing Talk. Part 2: The CLAN Program*. Carnegie Mellon University, https://doi.org/10.21415/T5G10R.

Malvern, D., Richards, B., Chipere, N. and Durán, P. 2004. *Lexical diversity and language development*. New York: Palgrave Macmillan.

McCarthy, P.M., Watanabe, S. and Lamkin, T.A. 2012. The Gramulator: A Tool to Identify Differential Linguistic Features of Correlative Text Types. In McCarthy, P.M. and Boonthum-Denecke, C., (eds.), *Applied Natural Language Processing: Identification, Investigation and Resolution*, 312-333. IGI Global.

McGillivray, B and Kilgarriff, A. 2013. Tools for Historical Corpus Research, and a Corpus of Latin. In P. Bennett, M. Durrell, S. Scheible and R.J. Whitt, (eds.), *Methods in Historical Corpus Linguistics*. Tübingen: Narr.

McGillivray, B., Hengchen, S., Lähteenoja, Palma, M. and Vatri, A. 2019. A computational approach to lexical polysemy in Ancient Greek. *Digital Scholarship in the Humanities*.

Meroño-Peñuela, A., De Boer, V., Van Erp, M., Melder, W., Mourits, R., Schalk, R. and Zijdeman, R. 2020. Ontologies in CLARIAH: Towards Interoperability in History. *Language and Media*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and J. Dean 2013. Distributed Representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L., Souma, W. and Trajanov, D. 2019. Forecasting corporate revenue by using deep-learning methodologies. In *2019 International Conference on Control, Artificial Intelligence, Robotics and Optimization ICCAIRO*, 115–120. IEEE.

Mosallanezhad, A., Beigi, G. and Liu, H. 2020. Deep reinforcement learning-based text anonymization against private-attribute inference. In *EMNLP-IJCNLP 2019 - 2019 Proceedings of the Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, 2360-2369.

OpenCV. 2017. *Introduction to Support Vector Machines — OpenCV 2*.

Othman, M.S., Hassan, H.A. and Moawad, R. 2014. Opinion mining and sentimental analysis approaches: A survey. *Life Science Journal*, 114, 321-326.

Pavelko, S.L., Owens, R.E., Ireland, M. and Hahs-Vaughn, D. L. 2016. Use of language sample analysis by school-based SLPs: Results of a nationwide survey. *Language, Speech and Hearing Services in Schools*, 47, 246–258.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, June 2018*, 1 Long Papers, 2227–2237. Association for Computational Linguistics.

Pezold, M.J., Imgrund, C.M. and Storkel, H.L. 2020. Using Computer Programs for Language Sample Analysis. *Language, Speech and Hearing Services in Schools*, 511, 103–114. doi:10.1044/2019_LSHSS-18-0148.

Polpinij, J. and Ghose, A.K. 2008. An Ontology-Based Sentiment Classification Methodology for Online Consumer Reviews. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 518-524, doi: 10.1109/WIIAT.2008.68.

Rokas, A., Rackevičienė, S. and Utka, A. 2020. Automatic extraction of Lithuanian cybersecurity terms using Deep Learning approaches. Zenodo.

Romary, L., Khemakhem, M., Khan, F., Bowers, J., Calzolari, N., George, M., Pet, M. and Bański, P. 2019. *LMF reloaded*. arXiv preprint arXiv:1906.02136.

Rosin, G.D. and Radinsky, K. 2019. Generating Timelines by Modeling Semantic Change. In *Proceedings of the 23rd Conference on Computational Natural Language Learning CoNLL*. 186–95. Hong Kong: Association for Computational Linguistics. https://doi.org/10.18653/v1/K19-1018.

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. 2019. Distilbert, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. In *Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 1–5.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H. and Tahmasebi, N. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation, Barcelona, Spain*. Association for Computational Linguistics.

Soares, M.A. and Parreiras, F.S. 2020. A literature review on question answering techniques, paradigms and systems. *J. King Saud University Computation Information Science*, 32, 635-646.

Sohangir, S., Petty, N. and Wang, D. 2018. Financial sentiment lexicon analysis. In *2018 IEEE 12th International Conference on Semantic ComputingICSC*, 286–289. IEEE.

Sohangir, S., Wang, D., Pomeranets, A. and Khoshgoftaar, T.M. 2018. Big data: Deep learning for financial sentiment analysis. *Journal of Big Data*, 5.1, 3.

Souma, W., Vodenska, I. and Aoyama, H. 2019. Enhanced news sentiment analysis using deep learning methods. *Journal of Computational Social Science*, 2.1, 33–46.

Spyns, P., and Odijk, J. (eds.). 2013. *Essential Speech and Language Technology for Dutch*. Berlin Heidelberg: Springer.

Su, H., Shen, X., Zhang, R., Sun, F., Hu, P., Niu, C. and Zhou, J. 2019. Improving Multi-turn Dialogue Modelling with Utterance ReWriter. ACL.

Tahmasebi, N. Borin, L. Jatowt, A. 2019. Survey of Computational Approaches to Lexical Semantic Change. ArXiv:1811.06278 [Cs], March 13, 2019.

Tai, K.S., Socher, R. and Manning, C.D. 2015. Improved semantic representations from tree-structured long short-term memory networks, arXiv preprint arXiv:1503.00075.

Tang, D., Qin, B. and Liu, T. 2015. Document modelling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 1422–1432.

Tsakalidis, A. and Liakata, M. 2020. Sequential Modelling of the Evolution of Word Representations for Semantic Change Detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing EMNLP*, 8485–8497. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.682.

Vatri, A., and McGillivray, B. 2018. The Diorisis Ancient Greek Corpus. *Research Data Journal for the Humanities and Social Sciences*, 31. 55-65.

Vidgen, B., and Derczynski, L. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS one*, 1512. e0243300.

Wang, G., Wang, T., Wang, B., Sambasivan, D., Zhang, Z., Zheng, H. and Zhao, B.Y. 2015. Crowds on Wall Street: Extracting value from collaborative investing platforms. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, 17–30.

Wang, S., Schlobach, S. and Klein, M. 2011. Concept Drift and How to Identify It. *Journal of Web Semantics*. *First Look*.

Welty, C., Fikes, R. and Makarios, S. 2006. A reusable ontology for fluents in OWL. In *FOIS*, 150, 226–236.

Wohlgenannt, G. and Minic, F. 2016. Using Word2vec to Build a Simple Ontology Learning System. *International Semantic Web Conference, 2016*.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E. 2016. Hierarchical Attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 1480–1489.

Zapilko, B., Harth, A. and Mathiak, B. 2011. Enriching and Analysing Statistics with Linked Open Data. In *Conference on New Techniques and Technologies for Statistics*.

Zhang, L., Wang, S. and Liu, B. 2018. Deep learning for sentiment analysis:A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8.4, e1253.

Zhang, X., Zhao, J. and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, 649–657.

Zhao, L., Li, L. and Zheng, X. 2020. A bert based sentiment analysis and key entity detection approach for online financial texts, arXiv preprint arXiv:2001.05326.