D4.1

Use Case Description and Requirements Elicitation. Working Group 4 “Use Cases and Applications”

Main authors: Sara Carvalho, Ilan Kernerman

29 April 2021

Project Acronym	NexusLinguarum
Project Title	European network for Web-centred linguistic data science
COST Action	18209
Starting Date	26 October 2019
Duration	48 months
Project Website	https://nexuslinguarum.eu/
Chair	Jorge Gracia
Main authors	Sara Carvalho and Ilan Kernerman
Contributors	Florentina Armaselu, Mariana Damova, Kristina Despot, Daniela Gifu, Gordana Hržica, Valentina Janev, Barbara Lewandowska-Tomaszczyk, Ana Luís, Petya Osenova, Sigita Rackevičienė, Marko Robnik-Šikonja, Dimitar Trajanov, Jouni Tuominen, Slavko Žitnik; [Julia Bosque Gil, Christian Chiarcos]
Reviewer	[Jorge Gracia]
Version | Status	V1.0 | final
Date	29 April 2021

Acronyms List

CA	COST Action
LD	Linked Data
LLD	Linguistic Linked Data
LLOD	Linguistic Linked Open Data
LOD	Linked Open Data
LR	language resource
NLP	Natural Language Processing
SA	Sentiment Analysis
SALLD	Sentiment Analysis and Linguistic Linked Data
SOTA	state-of-the-art
STSM	Short Term Scientific Mission
UC	Use Case
WG	Working Group
WSD	word sense disambiguation

Table of Contents

Executive Summary

Working Group 4 (WG4) of the NexusLinguarum COST Action is dedicated to use cases and applications where the Action's relevant methodologies and technologies can be tested and validated. This deliverable reports on the state of play as of M18 (April 2021) and focuses on detailed specifications of the various task domains and use cases (UCs), including the elicitation of the requirements necessary for their implementation. In addition, we describe the collaboration with other WGs and a related workshop. Finally, an outlook of further steps is presented, along with extensive references.

1. Introduction

Working Group 4 (WG4, Use cases and applications) works closely with the remaining NexusLinguarum WGs, providing, via a set of relevant use cases, a context for the practical application – and validation – of the technologies, methodologies, and standards developed within the scope of the Action. With approximately 110 members (as of February 22, 2021), WG4 integrates participants from nearly 40 countries and with various backgrounds, thereby fostering interdisciplinary collaboration. WG4's structure incorporates this approach and directly benefits from it by having two leaders for each Task, with backgrounds in Linguistics and Computer Science, respectively.

Following the kick-off and first Management Committee meetings, it was decided to restructure the list of intended Tasks included in the original Memorandum of Understanding (MoU) by removing the legal and policy domains and adding technology, as follows: Linguistics (T4.1), Humanities and Social Sciences (T4.2), Technology (T4.3), and Life Sciences (T4.4).

Given the broad spectrum of these domains, the WG also found it pertinent to explicitly integrate a second level into the structure, where the actual use cases and applications scenarios could be further developed and discussed.

As a result, in the initial phase of the Action (first 18 months), the four Tasks include the following Use Cases (UCs):

Task 4.1: Use cases in Linguistics

UC4.1.1: Use Case in Media and Social Media

UC4.1.2: Use Case in Language Acquisition

Task 4.2: Use cases in Humanities and Social Sciences

UC4.2.1: Use Case in Humanities

UC4.2.2: Use Case in Social Sciences

Task 4.3: Use cases in Technology

UC4.3.1: Use Case in Cybersecurity

UC4.3.2: Use Case in Fintech

Task 4.4: Use cases in Life Sciences

UC4.4.1: Use Case in Public Health

UC4.2.2: Use Case in Pharmacy

It should be noted, however, that this list is not closed, and other fields may be added at the request of the Action members or interested communities. In fact, in M17, a call has been launched for new UCs, but no additional proposals have been received so far.

This document thus outlines the Tasks and Use Cases currently integrated into the WG by providing a thorough description of their objectives, methodologies, resources, milestones, and expected deliverables due by the end of the Action. Then, it reports in detail on the specific requirements for each UC, namely in what concerns the tools and technologies necessary for its implementation. In addition, it describes the cooperation with the other WGs and presents the SALLD-1 workshop which has been initiated within WG4 and will be held at LDK 2021 (M23). Finally, it explains the next steps and provides an extensive list of relevant literature. It should be mentioned that until April 2021, several publications have already been prepared and submitted by the different UCs to international journals and conferences focusing on the core topics underlying NexusLinguarum (e.g., the Semantic Web Journal or the 3rd Language, Data and Knowledge Conference – LDK) and will thus be updated and accounted for in the upcoming deliverable, due in October 2021.

2. Tasks and Use Cases

2.1. Task 4.1. Use Cases in Linguistics

Task Leaders	Kristina Despot (linguistic), Slavko Žitnik (computational)
Use Cases	
UC 4.1.1	Media and Social Media
UC 4.1.2	Language Acquisition

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., in lexicography, terminology, typology, syntax, comparative linguistics, computational linguistics, corpus linguistics, phonology etc.). General tasks within this task include: SOTA for the usage of LOD in Linguistics; document describing requirements elicitation and use case definition (M18); intermediate and final activity reports (M24 and M48); scientific papers on in-use applications of LLOD, NLP and linguistic big data (M48).

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

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

Coordinator Barbara Lewandowska-Tomaszczyk

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, their annotation and offensive content extraction will be exemplified and analysed.

The State-of-the-Art

Data

- 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), 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 (Technologies)

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

Languages English, Croatian, Hebrew, Lithuanian, Montenegrin, Polish

Roadmap

- Survey/selection of corpora

An online workshop was held to discuss the computational aspects involved in each of the planned tasks (end 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

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 and 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 a more detailed description of abusive language event structure and relevant abusive language scales developed in the use case.

Tasks (and persons responsible)

The tasks will be conducted in parallel throughout the Action lifetime (details will be provided by the coordinator and collaborators)

T1. Description of details of the use case objectives and implementation: (abuse, implicitness/explicitness, emotions/sentiments, hate speech) select languages – (Barbara Lewandowska-Tomaszczyk, Olga Dontcheva-Navratilova, Marcin Trojszczak)

T2. Selection of English hate speech datasets for analysis (Milica Vuković Stamatović. Branka Zivukovic)

T3. Survey of accessible sets of abuse language dimensions (Ana Ostroski, Lobel Filipić)

T4. Identification of explicit vs implicit abuse identificatory and classification criteria – direct literal vs indirect and figurative (Kristina Depot, Marcin Trojszczak)

T5. Development of abusive language identification scales (Barbara Lewandowska-Tomaszczyk/Jelena Mitrovič, Olga Dontcheva-Navratilova, Marcin Trojszczak); TYPES of abuse/accompanying EMOTIONS (Sentiment analysis+) (Barbara Lewandowska-Tomaszczyk, Marcin Trojszczak, Paul Wilson)

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 baseline models (Slavko Zitnik, Giedre Valunaite Oleskevicienė, Lobel Filipić)

T7. Abusive language tagset enrichment proposals (Chaya Liebeskind, contribution from all other use case members)

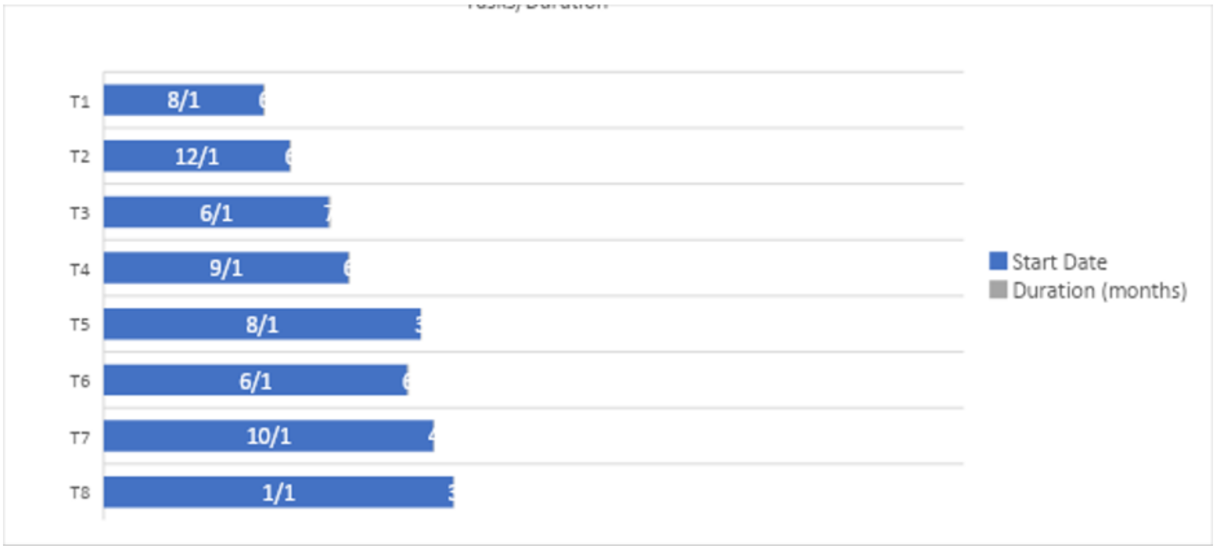
T8. Survey of LLOD infrastructure relevant to the task topic (ontologies of opinion mining, etc.). Infrastructure proposals of abusive hate speech data sharing (Slavko Zitnik, Barbara Lewandowska-Tomaszczyk)

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 winter-spring 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 and Methodology

Duration: 01.06.2020 – 1.10.2023



All Tasks start on 01.06.2020.

The methodology is divided into two interrelated groups, working in parallel.

- ***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.

NOTE. The data/tagging are planned to be selected within ca. 2 months to make it possible for the UC team members to select and prepare appropriate computational models and work in parallel.

- ***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 Science, WG4 UC4.3.1. Use Case in Cybersecurity, and others)
- STSMs
- Other (beyond Nexus, if appropriate) CLARIN, TRAC, LREC

Dissemination

- Reports
    - D1, D2, D3, D4, D5, D6, D7. Final Report
- Meetings, workshops and activities
- An international conference on *Explicit and Implicit Abuse* will be organized by UC 4.1.1 members in 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

2.1.2. UC 4.1.2. Use Case on Language Acquisition

Coordinator Gordana Hržica

Overview

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

A language sample (written or spoken text produced by 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, while for the second language acquisition they are usually written. This type of the 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, Owens, Ireland, & Hahs-Vaughn, 2016). During the last decades of the 20th century, computer programs were developed to assist language sample analysis (overview: Pezold, Imgurund, Storkel, 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 previously developed programs. Web-based programs usually concentrate on one domain. For example, the Gramulator tool (McCarthy, Watanabe & Lamkin, 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 allows for the implementation of measures such as lemma-token ratio, lexical density (content words/number of words) or similar.

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 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, Crossley, Berger, 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).

The State-of-the-Art

- **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) measures of cohesion.

- **Tools (Technologies)**

There are some existing computer programs used by the research community that are not user-friendly (CLAN – Computerized Language Analysis[1], SALT - Systematic Analysis of Language Transcripts[2]). 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).

Roadmap

- **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

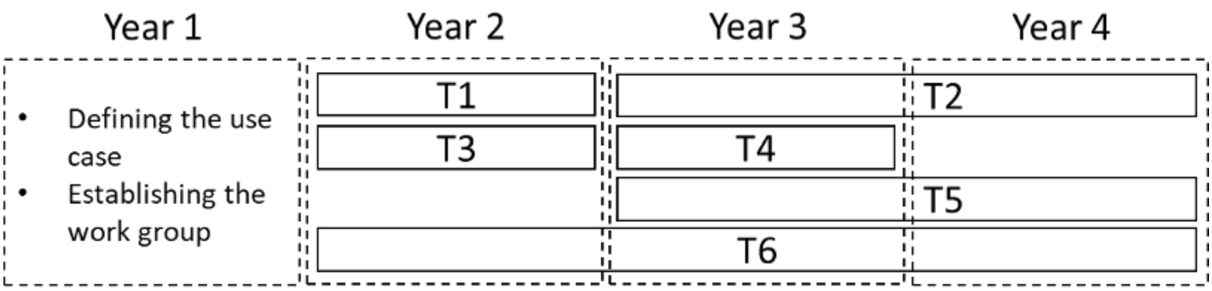
T3: Developing a survey for collecting information about the language sample analysis in individual countries

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 and Methodology



---

[1] https://dali.talkbank.org/clan/

[2] https://www.saltsoftware.com/

Methodology

extensive literature research, scientific networking, online survey

Deliverables

Year 1:
D1: WG 4 use case description for 4.1.2.
D2: NexusLinguarum use case template for 4.1.2.
Year 2:
D3: Overview of available language sampling tools
D4: Overview of language technologies for participating languages
Year 3:
D5: Survey for collecting information about the language sample analysis available online
D6: At least 50 researchers and/or practitioners have participated in the survey
D7: Results of the survey analysed and presented
Year 4:
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
- STSMs
- Other: speech and language pathologists, language teachers, computational linguists

Dissemination

- Reports
- Meetings, Workshops, Conferences
- Publications

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

Task Leaders Ana Luís (linguistics) (October 2019 - March 2021), Jouni Tuominen (computational)

Use Cases

UC 4.2.1	Humanities
UC 4.2.2	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. This task 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 will focus 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 will study 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.

2.2.1. UC 4.2.1. Use Case in Humanities

Coordinator Florentina Armaselu

Overview

The use case will focus on the evolution of parallel concepts in different languages and Humanities fields (history, literature, philosophy, religion, etc.). The methodology will include various textual collections and resources from corpus linguistics, word embedding and Linguistic Linked Open Data (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.

The State-of-the-Art

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 & Kilgarriff, 2013); Diorisis (7th century BC - 5th century AD) (McGillivray et al., 2019; Vatri & McGillivray, 2018); Responsa (11th century until now) (Liebeskind & 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 & 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 & Radinsky, 2019; Wohlgenannt & Minic 2016).
- Documenting, “explainable AI” (Hyvönen, 2020).

Tools (Technologies)

- 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 (Drymonaset al., 2010); Text2Onto (Cimiano & Volker, 2005).
- Semantic Web formalisms: RDF, OntoLex-Lemon, OWL-Time.
- SemEval 2020 task Unsupervised lexical semantic change detection (Schlechtweg et al., 2020).

Languages

- Ancient Greek, Hebrew, French, Latin, Old Lithuanian, other (TBD).

Roadmap

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 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 (and **persons** responsible)

Task	Description	Person responsible
T0	Define use case and participation.	Florentina Armaselu
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.	Florentina Armaselu Chaya Liebeskind Barbara McGillivray Giedrė Valūnaitė Oleškevičiene
T2	Draw the state-of-the-art (SOA) 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.	Florentina Armaselu Elena-Simona Apostol Anas Fahad Khan Chaya Liebeskind Barbara McGillivray Ciprian-Octavian Truică Andrius Utka Giedrė Valūnaitė Oleškevičiene Marieke van Erp
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 preprocessing (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, NER and semantic search engines.	Florentina Armaselu et al. Chaya Liebeskind Barbara McGillivray Giedrė Valūnaitė Oleškevičiene
T4	Study and choose the methods and tools for detecting semantic change and apply them to the selected data samples.	all
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.	all
T6	Document the whole process and produce a set of guidelines to describe the methodology derived from the use case.	Florentina Armaselu et al. Barbara McGillivray

Duration

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

Workflow and Methodology*

Task / Month	m6	m12	m18	m24	m30	m36	m42	m48
T0. Define use case and participation	⬜	⬜						
T1. Select concepts, languages		⬜	🟨					
T2. SOA. Model for concept change			🟨					
T3. Chose datasets and time spans			🟨	🟪				
T4. Study and apply change detection				🟪	🟪	🟪		
T5. Study and build LLOD ontologies					🟪	🟪	🟪	
T6. Document tasks, create guidelines	⬜	⬜	🟨	🟪	🟪	🟪	🟪	🟪

The methodology will involve 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.

* Colour codes: ⬜ - completed; 🟨 - in progress; 🟪 - not started.

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

- M1. Theoretical framework of the use case (m18).
- M2. Selected datasets to be processed (m24).
- M3. Change detection results (m36).
- M4. LLOD published ontologies (m42).
- M5. 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);
    - WG 3 - 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).
- STSMs
- Other (beyond Nexus, if appropriate)
    - Possible participation in the ADHO SIG-LOD.
    - Possibly applying for funding (e.g., European programme, if available, for an extended version of the use case).

Dissemination

- Reports
    - D1, D2, D3, D5.
- Meetings, Workshops
    - Nexus activities.
- Conferences, Publications
    - DH, LDK, LREC, COLING, ISWC, SEMANTiCS, Semantic Web conferences and journals.
- Submitted papers (under review):
    - *Semantic Web journal, Special Issue on Latest Advancements in Linguistic Linked Data:* LL(O)D and NLP Perspectives on Semantic Change for Humanities Research.
    - *LDK 2021 – 3rd Conference on Language, Data and Knowledge, 1-3 September in Zaragoza, Spain:* HISTORIAE, HIStory of culTural transfORmatIon as linguistIc dAta sciEnce. A Humanities Use Case.

2.2.2. UC 4.2.2. Use Case in Social Sciences

Coordinator Mariana Damova

Overview

Survey data provide a valuable source of information and research for different scientific disciplines, such as social sciences, philosophy, anthropology, political sciences, history. They are also of interest for practitioners such as policy makers, politicians, government bodies, educators, journalists, and all other stakeholders with occupations related to people and society. That is why 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 are data anonymization tools, semantic search, semantic data integration, relations detection. Further, 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.

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 answers 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 and concepts, and stemming
- **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).

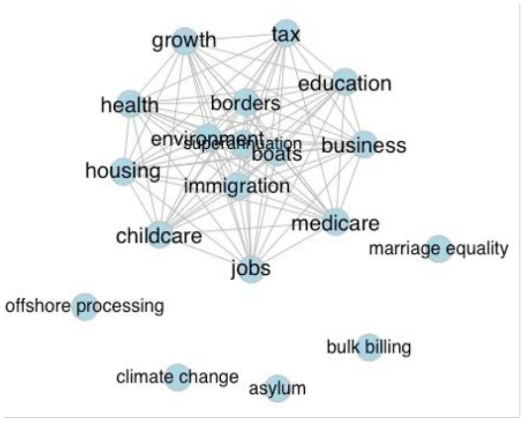


Figure 1

- **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, Hassan & Moawad, 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 of 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, Maas & Miles, 2002), and about religions (LICR)[3] 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, Harth & Mathiak, 2011). Finally, Linked Open Data Technologies are used to describe the catalogues of social data repositories, like in the CESSDA.eu catalogue[4]. 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

In the course of the project, we will use survey data and social media data.

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.

---

[3] https://datasets.iisg.amsterdam/dataset.xhtml?persistentId=hdl:10622/MHJWRZ

[4] https://datacatalogue.cessda.eu/

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, and 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 state-of-the-art 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 come up with workflows and guidelines for the adoption of language processing approaches depending on the datasets to be targeted. Further, we will elaborate workflows for datasets curation including data anonymisation techniques (Kleinberg et al., 2017; Mosallanezhad et al., 2019), and user profiling. Further, 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, as well as 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 & Parreiras, 2020) for better interpretation of the free text answers to open questions from the surveys and maybe full surveys.

Tools (Technologies)

Apart from the approaches listed in the state-of-the-art 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[5,6] 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, 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.

---

[5] https://www.linguisticanalysistools.org/sinlp.html

[6] https://www.linguisticanalysistools.org/tools.html#

Text analysis[7] uses Natural Language Processing (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[8] 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 to ultimately analyse the spread of memes and opinions through repeated interactions in linguistic communities.

MonkeyLearn[9] has a number of pre-trained models that can help you analyse your survey results right away. For example, our sentiment analysis model will help you see if your customers' responses are *Negative*, *Positive*, or *Neutral*, while our aspect classifier identifies the theme or topic those customers mention.

SPSS Analytics Partner[10] IBM SPSS Text Analytics for Surveys uses powerful natural language processing 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 you uncover patterns in the attitudes, beliefs and opinions of others.

Perceptix[11] 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**[12] 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.

---

[7] https://monkeylearn.com/blog/survey-analysis/

[8] https://www.oii.ox.ac.uk/study/courses/introduction-to-natural-language-processing-for-the-social-sciences/

[9] https://monkeylearn.com/

[10] https://www.spssanalyticspartner.com/software/ibm-spss-text-analytics-for-surveys/

[11] https://blog.perceptyx.com/open-ended-survey-questions-analysis

[12] https://www.european-language-grid.eu/

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

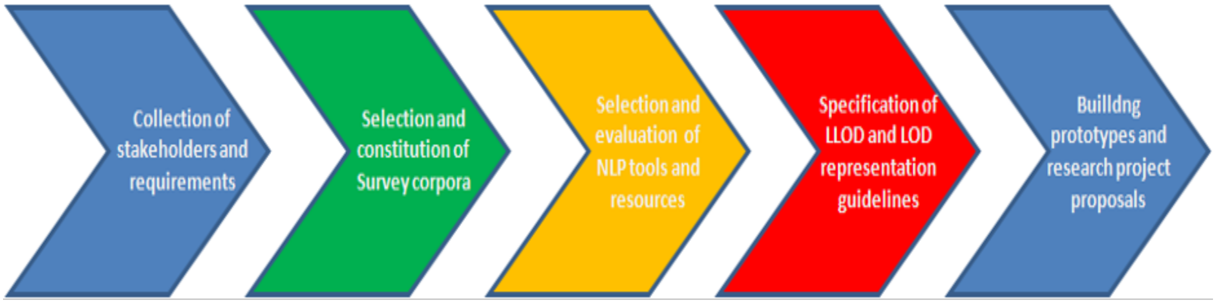

Collection of stakeholders and requirements
Selection and constitution of Survey corpora
Selection and evaluation of NLP tools and resources
Specification of LLOD and LOD representation guidelines
Builldng prototypes and research project proposals

Figure 2

Strategy

Our strategy has three pillars:

- 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

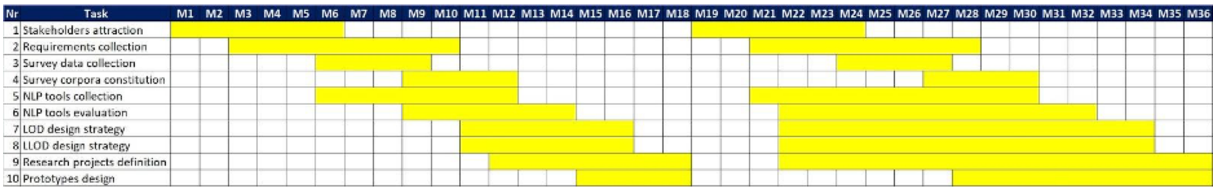
Tasks (and persons responsible)

Nr	Task	Description	Responsible
1	Stakeholders attraction	Identification, contact, awareness raising and attraction of stakeholders	Giedre Valunaite Oleskeviciene, Mariana Damova, et al.
2	Requirements collection	Interviewing stakeholders and definition of users, specification of requirements	Giedre Valunaite Oleskeviciene, Mariana Damova, Radovan Garabik et al.
3	Survey data collection	Collection of surveys corresponding to the topics of interest	Giedre Valunaite Oleskeviciene, Mariana Damova, Chaya Liebeskind et al.
4	Survey corpora constitution	Analysis of the collected surveys and constitution of corpora in easy for processing format	Mariana Damova, Chaya Liebeskind
5	NLP tools collection	Selection of NLP tools corresponding to the topics of interest and to the requirements	Mariana Damova, Dimitar Trajanov, Dagmar Gromann
6	NLP tools evaluation	Evaluation of the selected NLP tools	Mariana Damova, Dimitar Trajanov, Chaya Liebskind, Dagmar Groman
7	LOD design strategy	Analysis and definition of the adoption of LOD for Survey processing based on the defined requirements	Jouni Tuominen, Mariana Damova, Eveline Wandl-Vogt, Chirstian Chiarcos
8	LLOD design strategy	Analysis and definition of the adoption of LLOD for Survey processing based on the defined requirements	Jouni Tuominen, Mariana Damova, Eveline Wandl-Vogt, Christian Chiarcos et al.
9	Research projects definition	Definition of research topics, formation of project consortia, submission of research proposals	all
10	Prototypes design	Description of guidelines for developing resources, tools or solutions for surveys processing with LLOD and LOD methods	all

Duration

From July 1, 2020 – June 2023 (36 months)

Workflow and Methodology



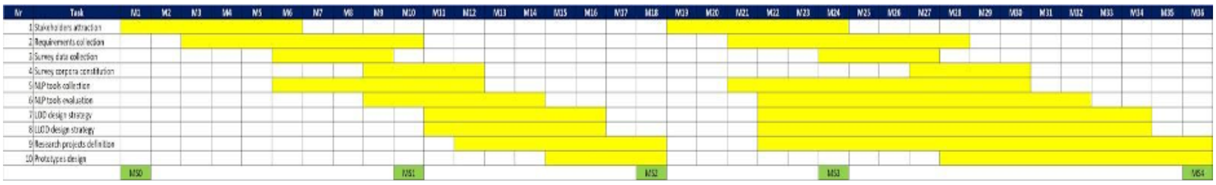
Nr	Task	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22	M23	M24	M25	M26	M27	M28	M29	M30	M31	M32	M33	M34	M35	M36
1	Stakeholders attraction	■	■	■	■	■	■													■	■	■	■	■	■												
2	Requirements collection			■	■	■	■	■	■	■	■											■	■	■	■	■	■	■	■								
3	Survey data collection						■	■	■	■															■	■	■	■									
4	Survey corpora constitution									■	■	■	■															■	■	■	■						
5	NLP tools collection						■	■	■	■	■	■	■									■	■	■	■	■	■	■	■	■	■						
6	NLP tools evaluation									■	■	■	■	■	■								■	■	■	■	■	■	■	■	■	■	■				
7	LOD design strategy											■	■	■	■	■	■						■	■	■	■	■	■	■	■	■	■	■	■	■		
8	LLOD design strategy											■	■	■	■	■	■						■	■	■	■	■	■	■	■	■	■	■	■	■		
9	Research projects definition												■	■	■	■	■	■	■				■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
10	Prototypes design															■	■	■	■										■	■	■	■	■	■	■	■	■

Deliverables

All deliverables will report about all 10 tasks

1. Initial Use case design - M18
2. Intermediary Use case design - M24
3. Final Use case design - M36

Milestones



Nr	Task	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22	M23	M24	M25	M26	M27	M28	M29	M30	M31	M32	M33	M34	M35	M36
1	Stakeholders attraction	■	■	■	■	■	■													■	■	■	■	■	■												
2	Requirements collection			■	■	■	■	■	■	■	■											■	■	■	■	■	■	■	■								
3	Survey data collection						■	■	■	■															■	■	■	■									
4	Survey corpora constitution									■	■	■	■															■	■	■	■						
5	NLP tools collection						■	■	■	■	■	■	■									■	■	■	■	■	■	■	■	■	■						
6	NLP tools evaluation									■	■	■	■	■	■								■	■	■	■	■	■	■	■	■	■	■				
7	LOD design strategy											■	■	■	■	■	■						■	■	■	■	■	■	■	■	■	■	■	■	■		
8	LLOD design strategy											■	■	■	■	■	■						■	■	■	■	■	■	■	■	■	■	■	■	■		
9	Research projects definition												■	■	■	■	■	■	■				■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
10	Prototypes design															■	■	■	■										■	■	■	■	■	■	■	■	■
		MS0									MS1								MS2						MS3												MS4

Collaboration and Exchange

- UC coordination and WG4 communication channels - Slack
- WG4 UC 4.1.1
- Nexus WG1, WG2
- STSMs (TBD)
- CESSDA, FORSCENTER, University of Gothenburg, Bulgarian Academy of Sciences and others
- Consortia for H2020 and other bilateral funding for R&D opportunities
- Slack own channel for WG4 Use case Social Sciences
- Bi-weekly meeting on Tuesday at 11:00 PM CET

Dissemination

- Reports
- Meetings, Workshops, Conferences
- Publications

2.3. Task 4.3. Use Cases in Technology

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

Use Cases

UC 4.3.1	Cybersecurity
UC 4.3.2	Fintech

Overview

Task 4.3 builds upon the recent advancements in the areas of multilingual technologies, machine translation, automatic term extraction methods, text analytics and sentiment analysis models with the aim to reuse existing open-source components and test them in different ICT and business scenarios. General subtasks within this task include: state-of-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 of the CA, 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 the COST Action will impact involved researchers and industrial users of language technologies.

2.3.1. UC 4.3.1. Use Case in Cybersecurity

Coordinator: Sigita Rackevičienė

Participants: Liudmila Mockienė, Andrius Utka, Aivaras Rokas, Valentina Janev

Overview

The aim of the use case: to develop a methodology for the compilation of a termbase for under-resourced languages using deep learning systems and LLOD principles, in addition to applying it to the domain of cybersecurity (CS). The datasets will encompass both parallel and comparable corpora, which will provide 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. The state-of-the-art neural networks will be developed and applied for automatic extraction of terminological data and metadata necessary for termbase compilation.

The State-of-the-Art

Resources for datasets

- 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

- dataset collection methodology: compilation of parallel and comparable corpora; development of manually annotated gold standard corpora;
- automatic term extraction and alignment methodology: development and application of deep learning systems using gold standard datasets as training data;
- knowledge-rich context extraction methodology: development and application of knowledge-rich context extraction methods;
- development of an interlinked termbase using LLOD.

Tools (Technologies to be developed under the use case): manual annotation software, neural networks for automatic data extraction.

Languages: English and Lithuanian.

Roadmap

Strategy

The strategy implies developing a methodology for terminology management targeted at under-resourced languages, which would enhance the quality and reusability of termbases.

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

Tasks

T1: Research on existing cybersecurity terminology: searching and getting acquainted with existing English cybersecurity termbases, glossaries, ontologies; systematisation of the collected information (Sigita, Liudmila, Andrius).

T2: Compilation of corpora - building a knowledge store (Sigita, Liudmila, Andrius and consulting CS specialists):

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; security policy documents of social media portals, etc.)

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

T2.3. Development of gold standard corpora with manually labelled CS terms for training and assessment of machine learning and neural network systems.

T3: Automatic extraction of terminological data and metadata (Andrius and consulting CS specialists):

T3.1. Iterative testing of the automatic term extraction methods by comparing their results with the gold standards;

T3.2. Selection of the most effective methods and automatic extraction of term candidates from parallel and comparable corpora, their automatic alignment;

T3.3. Selection of the dominant CS terms based on frequency/dispersion analysis and expert approval.

T3.4. Development of automatic methods for extraction of knowledge-rich contexts; their extraction for the selected dominant CS terms.

T4: Compilation of the termbase (Sigita, Liudmila, Andrius, consulting CS specialists):

T4.1. Formulating final definitions of the terms using the extracted knowledge-rich contexts;

T4.2. Collecting other metadata about the selected terms: usage examples; conceptual relations with other terms, statistical data on term frequency and dispersion, etc.

T4.3. Uploading the collected data to a termbase.

T5: Interlinking the termbase· with other resources and its application (Aivaras, Andrius, Sigita, Liudmila, Valentina) Interlinking the termbase with other resources and its application in the cybersecurity domain (cross-lingual retrieval).

T6: Analysis of the conceptual and linguistic dimensions of the collected terminology (Sigita, Liudmila, Andrius).

Workflow

Duration: June 2020 - October 2023

T0: May-June 2020	M0: Use case template development
T1: June 2020 – August 2020,	M1: Description of SOTA
T2: June 2020 - June 2021	M2: Compilation of the corpora, development of gold standard corpora
T3: July 2021 - December 2021	M3: Automatic data extraction methodology development and application
T4: January 2022 - June 2022	M4: Compilation of the termbase
T5: July 2022 - December 2023	M5: LLOD application
T6: January 2023 - October 2023	M6: Terminology analysis

Deliverables

D1: Parallel and comparable corpora of the CS domain. The corpora will be made available to the public in the CLARIN repository.

D2: Termbase of CS terms. The base will be publicly available on the internet. The compiled termbase could be used as a model for development of termbases, applying automatic term extraction methods in other domains and other languages. These technologies are especially relevant for under-resourced languages which lag behind in their development.

D3: Publications on the results of the use case.

Collaboration and Exchange

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

Dissemination

- Reports, meetings, workshops, conferences, publications

Publication within the scope of the UC

Rokas, A., Rackevičienė, S. & Utka, A. (2020). Automatic extraction of Lithuanian cybersecurity terms using Deep Learning approaches. Zenodo. https://doi.org/10.3233/FAIA200600

2.3.2. UC 4.3.2. Use Case in Fintech

Coordinator Dimitar Trajanov

Overview

The financial systems are one of the most dynamic and innovative systems in the world. 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 natural language processing (NLP) techniques.

Sentiment analysis 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. Sentiment analysis models can provide an efficient method for extracting actionable signals from the news. General sentiment analysis 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 Finance 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 out of which 10 are numeric features and 10 are categorical features, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014] (see https://archive.ics.uci.edu/ml/datasets/Bank+Marketing)

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 (Technologies)

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

Languages

- English, …

Roadmap

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 and Methodology

	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

- M1. Survey of the current approaches
- M2. Evaluation of the models
- M3. 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

Conferences (selection)

(1) BPM 2020: 18th International Conference on Business Process Management, Sevilla, Spain, September 15-17, 2020 (https://congreso.us.es/bpm2020/) - Rank A

(2) 17th The IEEE International Conference on e-Business Engineering (ICEBE), Guangzhou, China, October 16-18, 2020 (https://conferences.computer.org/icebe/2020/index.htm) - Rank B

Journals (selection)

(1) Business & Information Systems Engineering (impact factor = 3.6) - https://www.springer.com/journal/12599/

(2) Business Process Management Journal (rank B on CORE Platform; impact factor = 1.46) - https://www.emeraldgrouppublishing.com/journal/bpmj

(3) Business Intelligence Journal (rank C_CORE Platform) - https://tdwi.org/research/list/tdwi-business-intelligence-journal.aspx

2.4. Task 4.4. Use Cases in Life Sciences

Task Leaders Petya Osenova (linguistics) (October 2019 - April 2021) and Marko Robnik-Šikonja (computation)

Overview

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

The State-of-the-Art

- **Resources**

We will rely on several types of resources: available Ontologies, Corpora and Lexical databases (such as Terminological dictionaries)

- **Methods**

When data is identified and gathered, as well as the related ontologies and lexicons, the following methods will be applied: Machine Learning, Information Extraction, and NLP. The linguistic pipelines such as Stanza covers most of European languages and provide the baseline text processing, such as tokenization, lemmatization, POS-tagging and to a lesser degree dependency parsing.

- **Technologies and Approaches**

The approaches include: Linked Open Data, Embeddings, Knowledge Graphs.

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

- **Languages**

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

Roadmap

- **Strategy**

  We will start 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 (and persons responsible)**
  1. State-of-the art Overview
     - General trend in life sciences (Petya, Marko, Slavko, Eveline, Dimitar, Konstantinos, Sara, Daniela, Ana)
     - In Public Health and Ecology (Konstantinos, Ana)
     - In Pharmacy (Dimitar, Marko)
  2. Identification of the related resources and tools
     - Resources (Sara, Daniela, Eveline, Petya, Ana)
     - Tools (Slavko, Marko)
  3. Description of their status (advantages, problems, etc.)
     - Resources (Petya and linguists)
     - Tools (Marko and CSists)
  4. Preparation of datasets (All)
  5. Analytics
     - Information retrieval (Slavko)
     - Knowledge extraction (Dimitar, Konstantinos)
     - Explanation of models (Marko)

- **Duration**
  1. July 2020 (M1)
  2. December 2020 (M6) – January 2021 (M7)
  3. June 2021 (M12)
  4. July 2021 (M13) – January 2022 (M19)
  5. January 2022 (M19) – April 2022 (M22)
  6. May 2022 (M23) – October 2023 (M28)

Workflow and Methodology

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

Deliverables

- Deliverable on SOTA (February 2021 (M8))
- Deliverable on identification and description of related resources and tools (January 2022 (M19))
- Deliverable on datasets and analytics (October 2023 (M28))

Milestones

- **M1:** SOTA - January 2021 (M7)
- **M2**: Identification of LRE – June 2021 (M12)
- **M3:** Description of available LRE – January 2022 (M19)
- **M4:** Datasets – April 2022 (M22)
- **M5:** Probes on Analytics Approaches – October 2023 (M28)

Collaboration and Exchange

- UC coordination and WG4 communication channels:
  - Emails, Google Drive, virtual meetings
- WG4 UCs and Tasks
  - UC4.1.1: expertise in working with social media
  - T4.3: expertise with regards to technology
- Nexus WGs
  - Interaction with all other WGs
- STSMs
  - At the moment no STSMs are planned given the complex international situation due to COVID-19.
- Other (beyond Nexus, if appropriate)

Dissemination

- Reports
    - All planned deliverables will serve also as reports
- Meetings, Workshops, Conferences
    - We will meet at the annual COST meeting.
    - If there is interest, we can organize a domain-related workshop.

Links about initiatives on COVID-19 data

EU Open Data Portal: https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data

Novel Coronavirus (COVID-19) Cases Data: https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases

5 Datasets About COVID-19: https://towardsdatascience.com/5-datasets-about-covid-19-you-can-use-right-now-46307b1406a

BioPortal: A dataset of linked biomedical ontologies:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4159173/

WikiData on COVID-19: https://www.wikidata.org/wiki/Q84263196

International Statistical Classification of Diseases and Related Health Problems (but: in PDF): (here in Bulgarian):
https://srzi.bg/uploads/pages/Lechebni_zavedeniq/3.MKB_10/1_mkb_v1_part1.pdf

NOTE: In March 2021, it was decided to form a new Use Case in Public Health (UC4.4.1) to handle the crux of this task, as described above, and is coordinated by the two task leaders. In addition, a new Use Case in Pharmacy (UC 4.4.2) was formed and is coordinated by Dimitar Trajanov.

3. Requirements' elicitation

This section outlines the specific requirements for the viability of each Use Case, as regards language resources, methodologies, technologies and tools, both on its own and in connection to other Tasks and UCs in WG4, as well as to other WGs. To gather this data, an iterative approach was adopted, starting with internal feedback within WG4's core group and, subsequently, the whole WG, followed by feedback from the other WG leaders and their respective core groups.

Overall, bullet points (•) refer to *internal* requirements, while squares (▢) elicit potential (or ongoing) contributions from the *other WGs* and their respective tasks.

UC4.1.1

- 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

UC4.1.2

- 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

UC4.2.1

- 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.)

UC4.2.2

- ❑ 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

UC4.3.1

- Compilation of parallel and comparable bilingual corpora of cybersecurity domain and of cybersecurity 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

UC4.3.2

- 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

T4.4

❑ 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)

4. Related activities

4.1. 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 remaining Nexus WGs, this has implied building up different forms of interaction between WG4 and the other WGs in general and between the WG4 UCs and other WG Tasks in particular. At the time of this Deliverable, the ties are mainly in initial stages. Here we provide a first overview of the current status.

4.1.1 List of Tasks of the other WGs and WG4 UCs

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
T2.5	Terminology & Knowledge 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
UC4.4.1	Public Health
UC4.4.2	Pharmacy

4.1.2 Table of Task/UC interaction

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

4.1.3 WG1 Tasks issues for collaboration with WG4

T1.1	modelling of corpus
T1.1	annotation of resources
T1.1	representation of diachronic information
T1.2	resource changes across time/different versions/gradual enrichment
T1.3	interlinking aspects
T1.4	(meta-)data quality aspects/assessment
T1.5+T1.2	resource development/transformation/reuse for under-resourced languages

4.1.4 Common ground between the other WG tasks and WG4 UCs

4.2.1-1.1-3.4	diachronic development in OntoLex + language tags for historical language stages (beyond OntoLex)
4.2.1-1.1-1.3-3.2	modelling of diachronic info; modelling dictionaries
4.2.1-1.2+2.1+2.5	generating and publishing multilingual parallel LLOD ontologies to trace the evolution of concepts
4.2.1-1.3+3.4	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
4.2.1-3.2	embeddings & deep learning for humanities (cf. MacBerth project)
4.2.1-3.5	providing examples of applications, combining LLOD and diachronic analysis, that may be used in teaching linguistic data science
4.2.2-1.1	modelling corpora & annotating resources (wrt FrAC, LingAnno within OntoLex)
4.2.2-3.2	collaboration on survey creation & (automated) evaluation

4.2 SALLD-1 Workshop

The first workshop on Sentiment Analysis and Linguistic Linked Data (SALLD-1) has been initiated in the context of WG4, in view of the Sentiment Analysis (SA) factor in relation to Linguistic Linked Data (LLD) as recurring in several Ucs. It was accepted for the third conference on Language, Data and Knowledge (LDK 2021, http://2021.ldk-conf.org/), in Zaragoza Spain, and is due to be held as a half-day pre-conference workshop on September 1, 2021.

The focus of SALLD-1 is on approaches that combine SA and LLD, which to our knowledge has not been undertaken (explicitly) before, in the aim of exploring relevant principles, methodologies, resources, tools and applications. It sets to present diverse perspectives on this joint subject matter, such as with regard to any domain (e.g. general media or social media, literary texts and digital humanities, fintech, cyber security, and so on); modelling and technical features; lexical resources and complex multi-level structure; emotion, hate speech and common words that become abusive in specific context; keywords, tags, polarity, standards, and any other related issue, including reviews or encodings and interconnection of SA tasks applying semantic technologies with LLD.

By the deadline of 23 April 2021, 10 submissions have been received and the review process has begun immediately afterwards.

Full details on the objectives, organisers and program committee, as well as regular updates, are available on the workshop website: https://salld.org/.

Concluding remarks and next steps

This report 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 integrating 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 STSMs could be implemented to support these collaborative interactions, which are, ultimately, one of the axes underpinning COST Actions.

References (selection)

Abirami, M.A.M. & 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, vol. 3, pp. 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," in Semantic Web Evaluation Challenge, pp. 124–128, Springer.

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

Betti, A. Van den Berg, H. (2014). "Modelling the History of Ideas". British Journal for the History of Philosophy 22, no. 4 (July 4, 2014): 812–35. https://doi.org/10.1080/09608788.2014.949217.

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

Buitelaar, P. Cimiano, P. and Magnini, B. (2005). "Ontology learning from text: An overview." In Ontology Learning from Text: Methods, Evaluation and Applications, volume 123, pages 3–12. IOS Press, 2005.

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, vol. 10318, 89–103, Cham, Switzerland, June. Springer.

Cimiano P., Chiarcos C., McCrae J.P., Gracia J. (2020). "Linguistic Linked Data in Digital Humanities." In Linguistic Linked Data, Springer, Cham. 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." 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. Montoyo A, Munoz R, Metais E (Eds); Springer: 227-238.

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

Crossley, S. A., Roscoe, R., Graesser, A., & 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, & A. Mitrovic (Eds.), Proceedings of the 15th International Conference on Artificial Intelligence in Education. (pp. 438-440). Auckland, New Zealand: AIED.

Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D.S. (2014). Analyzing discourse processing using a simple natural language processing tool (SiNLP). Discourse Processes, 51(5-6), pp. 511-534, DOI: 10.1080/0163853X.2014.910723

Davidson, T., Warmsley, D., Macy, M. W. & 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, pages 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, page 277–287. Springer Berlin Heidelberg. 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, page 7212–7221. European Language Resources Association (ELRA), licensed under CC-BY-NC, May 2020.

Fergadiotis, G., Wright, H. & Green, S. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. Journal of Speech, Language, and Hearing Research 58(3), 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. https://www.semanticscholar.org/paper/On-the-Semantics-of-Concept-Drift%3A-Towards-Formal-Fokkens-Braake/2ab391204c1e5397a6c50c71112c0520e29d6750.

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." J. Lang. Technol. Comput. Linguistics, 27(2):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, vol. 40, no. 16, pp. 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, pages 3960–3973, Online, July 2020. 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," in IEEE Intelligent Systems, vol. 17, no. 1, pp. 54-60, Jan.-Feb. 2002, 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, pages 1–11, Online, Nov. 2020. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/2020.conll-1.1.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & 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, volume 2, Apr 2010.

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, IOS Press, 2020. 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. https://www.semanticscholar.org/paper/A-Survey-on-Ontology-Enrichment-from-Text-Iyer-Mohan/.

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 (Volume 1: Long Papers),pp. 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.” Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020), Maxim Ionov, John P. McCrae, Christian Chiarcos, Thierry Declerck, Julia Bosque-Gil, Jorge Gracia (eds.), LREC 2020 Workshop, Language Resources and Evaluation Conference, 11–16 May 2020. https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/LDL2020book.pdf.

Kim, Y. (2014). “Convolutional neural networks for sentence classification,” arXiv preprint arXiv:1408.5882.

Kleinberg, B., Mozes, M., van der Toolen, Y., & Verschuere, B. (2018, January 31). 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, no. 3 (October 2008): 351–72. https://doi.org/10.1111/j.1468-2303.2008.00459.x.

Kutuzov, A. Øvrelid, L. Szymanski, T. Velldal, E. (2018). “Diachronic Word Embeddings and Semantic Shifts: A Survey.” Proceedings of the 27th International Conference on Computational Linguistics, Pages 1384–1397 Santa Fe, New Mexico, USA, August 20, 2018. https://www.aclweb.org/anthology/C18-1117/.

Kyle, K., Crossley, S., 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, vol. 15, no. 2, pp. 227–252.

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

Loughran, T. & McDonald, B. (2011). “When is a liability not a liability? textual analysis, dictionaries, and 10-ks,” The Journal of Finance, vol. 66, no. 1,pp. 35–65.

Luminoso, “Employee Feedback and Artificial Intelligence: A guide to using AI to understand employee engagement” (PDF file), downloaded from Luminoso website, https://luminoso.com/writable/files/White-Paper-Employee-Feedback-and-AI.pdf, 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. & Durán, P. (2004). Lexical diversity and language development. New York: Palgrave Macmillan.

McCarthy, P. M., Watanabe, S., & Lamkin, T. A. (2012). The Gramulator: A Tool to Identify Differential Linguistic Features of Correlative Text Types. In McCarthy, P. M., & Boonthum-Denecke, C. (Ed.), Applied Natural Language Processing: Identification, Investigation and Resolution (pp. 312-333). IGI Global. http://doi:10.4018/978-1-60960-741-8.ch01

McGillivray, B. Kilgarriff, A. (2013). "Tools for Historical Corpus Research, and a Corpus of Latin". In Methods in Historical Corpus Linguistics, Paul Bennett, Martin Durrell, Silke Scheible, Richard J. Whitt (eds.), Narr, Tübingen. https://www.researchgate.net/publication/236857134_Tools_for_historical_corpus_research_and_a_corpus_of_Latin.

McGillivray, B., Hengchen, S., Lähteenoja, Palma, M., Vatri, A. (2019). A computational approach to lexical polysemy in Ancient Greek, Digital Scholarship in the Humanities https://doi.org/10.1093/llc/fqz036.

Meroño-Peñuela, A. De Boer, V. Van Erp, M. Melder, W. Mourits, R. Schalk, R. Zijdeman, R. (2020). "Ontologies in CLARIAH: Towards Interoperability in History, Language and Media." https://Arxiv.Org/Abs/2004.02845v2.

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, pp. 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 & Optimization (ICCAIRO), pp. 115–120, IEEE.

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

OpenCV (2017). Introduction to Support Vector Machines — OpenCV 2. https://docs.opencv.org/3.4/d1/d73/tutorial_introduction_to_svm.html

Othman, M. S., Hassan, H. A., Moawad, R. (2014). Opinion mining and sentimental analysis approaches: A survey, In: Life Science Journal, 11(4), 321-326.

Pavelko, S. L., Owens, R. E., Ireland, M., & 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, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/N18-1202, doi:10.18653/v1/N18-1202.

Pezold, M. J., Imgrund, C. M., & Storkel, H. L. (2020). Using Computer Programs for Language Sample Analysis. Language, Speech & Hearing Services in Schools, 51(1), 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," 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 518-524, doi: 10.1109/WIIAT.2008.68.

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. 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, China: Association for Computational Linguistics, 2019. 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, pages 1–5.

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

Soares, M.A., & Parreiras, F.S. (2020). A literature review on question answering techniques, paradigms and systems. J. King Saud Univ. Comput. Inf. Sci., 32, 635-646.

Sohangir, S., Petty, N., and Wang, D. (2018). “Financial sentiment lexicon analysis,”in 2018 IEEE 12th International Conference on Semantic Computing(ICSC), pp. 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, vol. 5, no. 1, p. 3.

Souma, W., Vodenska, I. and Aoyama, H. (2019). “Enhanced news sentiment analysis using deep learning methods,” Journal of Computational Social Science, vol. 2, no. 1, pp. 33–46.

Spyns, P., & Odijk, J. (Eds.) (2013). Essential Speech and Language Technology for Dutch. Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30910-6

Su, H., Shen, X., Zhang, R., Sun, F., Hu, P., Niu, C., & 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. http://arxiv.org/abs/1811.06278.

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 modeling with gated recurrent neural network for sentiment classification,” in Proceedings of the 2015 conference on empirical methods in natural language processing,pp. 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), pages 8485–8497. Association for Computational Linguistics, Nov. 2020. doi:10.18653/v1/2020.emnlp-main.682.

Vatri, A., & McGillivray, B. (2018). The Diorisis Ancient Greek Corpus, Research Data Journal for the Humanities and Social Sciences, 3(1), 55-65. doi: https://doi.org/10.1163/24523666-01000013, https://brill.com/view/journals/rdj/3/1/article-p55_55.xml.

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. PloS one, 15(12), e0243300. https://doi.org/10.1371/journal.pone.0243300

Wang, G., T. Wang, B. Wang, D. Sambasivan, Z. Zhang, H. Zheng, and B. Y. Zhao (2015). “Crowds on wall street: Extracting value from collaborative investing platforms,” in Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 17–30.

Wang, S. Schlobach, S. Klein, M. (2011). “Concept Drift and How to Identify It”. Journal of Web Semantics First Look, September 2011. http://dx.doi.org/10.2139/ssrn.3199520.

Welty, C. Fikes, R. and Makarios, S. (2006). “A reusable ontology for fluents in OWL.” In FOIS, volume 150, pages 226–236.

Wohlgenannt, G. Minic, F. (2016). “Using Word2vec to Build a Simple Ontology Learning System.” International Semantic Web Conference, 2016. http://ceur-ws.org/Vol-1690/paper37.pdf.

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, pp. 1480–1489.

Zapilko, B., Harth, A., & Mathiak, B. (2011). Enriching and Analysing Statistics with Linked Open Data.

Zhang, L., Wang, S. and Liu, B. (2018). “Deep learning for sentiment analysis:A survey,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, p. e1253.

Zhang, X., Zhao, J. and LeCun, Y. (2015). “Character-level convolutional networks for text classification,” in Advances in neural information processing systems, pp. 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.