HISTORIAE, History of Socio-Cultural Transformation as Linguistic Data Science. A Humanities Use Case

Florentina Armaselu[1]
Centre for Contemporary and Digital History (C^{2}DH), University of Luxembourg, Luxembourg

Elena-Simona Apostol
Department of Computer Science and Engineering, Faculty of Automatic Control and Computer, University Politehnica of Bucharest, Romania

Anas Fahad Khan
Institute for Computational Linguistics «A. Zampolli»,
National Research Council of Italy, Pisa, Italy

Chaya Liebeskind
Department of Computer Science, Jerusalem College of Technology, Israel

Barbara McGillivray
Theoretical and Applied Linguistics, Faculty of Modern and Medieval Languages and Linguistics, University of Cambridge, UK
The Alan Turing Institute, London, UK

Ciprian-Octavian Truică
Department of Computer Science and Engineering, Faculty of Automatic Control and Computer, University Politehnica of Bucharest, Romania

Giedrė Valūnaitė Oleškevičienė
Institute of Humanities, Mykolas Romeris University, Vilnius, Lietuva

Abstract

The paper proposes an interdisciplinary approach including methods from disciplines such as history of concepts, linguistics, natural language processing (NLP) and Semantic Web, to create a comparative framework for detecting semantic change in multilingual historical corpora and generating diachronic ontologies as linguistic linked open data (LLOD). Initiated as a use case (UC4.2.1) within the COST Action *Nexus Linguarum, European network for Web-centred linguistic data science*, the study will explore emerging trends in knowledge extraction, analysis and representation from linguistic data science, and apply the devised methodology to datasets in the humanities to trace the evolution of concepts from the domain of socio-cultural transformation. The paper will describe the main elements of the methodological framework and preliminary planning of the intended workflow.

2012 ACM Subject Classification Computing methodologies → Semantic networks; Computing methodologies → Ontology engineering; Computing methodologies → Temporal reasoning; Computing methodologies → Lexical semantics; Computing methodologies → Language resources; Computing methodologies → Information extraction

Keywords and phrases linguistic linked open data, natural language processing, semantic change, diachronic ontologies, digital humanities

Digital Object Identifier 10.4230/OASIcs.LDK.2021.34

Author Contributions F.A., Sections 1, 2.1, 2.2, 2.5, 2.6, 2.7, 3; E.S.A., Section 2.4; A.F.K., Section 2.3; C.L., Section 1.3; B.M., Sections 1.3, 2.4; C.O.T., Section 2.4; G.V.O., Sections 1.3, 1.4. All the authors critically revised and approved the final version submitted to the LDK 2021 proceedings.

[1] florentina.armaselu@uni.lu

3rd Conference on Language, Data and Knowledge (LDK 2021).
Editors: Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch; Article No. 34; pp. 34:1–34:13

OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Use case description

1.1 Contextualisation

Semantic change has been studied so far within various disciplines and research fields, including the history of concepts and philosophy, linguistics, natural language processing (NLP) and the Semantic Web. Despite growing interest in the topic, which requires multiple perspectives and an interdisciplinary approach, there is no unified view and not enough dialogue on the subject, and different disciplines seem to make use of different interpretations and theoretical notions when dealing with it. Our proposal, called *HIstory of Socio-culTural transfORmation as linguistIc dAta sciEnce* (HISTORIAE) with reference to Tacitus's *Historiae* and nowadays interconnected cloud of linked data, aims at bridging the gap and combining approaches from these fields to create a comparative methodological framework for detecting semantic change in multilingual text collections and for generating corpus-based diachronic ontologies as linguistic linked open data (LLOD). The area of application of this proposal spans the digital humanities (DH), with a focus on the history of socio-cultural transformation in Europe and other regions, and emerging trends in knowledge extraction, analysis and representation from linguistic data science. These directions are noteworthy for current research, given the increasing use of digital and Web technologies in almost all the sectors of human activity and the need for a better understanding of their impact on cultural assets, within a broader historical, technological and data-aware context. It is expected that the project outcomes may also be applied to other domains.

HISTORIAE will address the following research questions. (1) Which insights does the study of semantic change help generate in the history of socio-cultural transformation? (2) Can the applied methodology inform us about the interrelation between linguistic, social and cultural innovation over time, and the socio-cultural roots of innovation? (3) What may be learned about the combination of human and machine agency in the process of construction and dissemination of knowledge, and of explaining the underlying mechanisms?

Throughout this paper, the term "semantic change" will generally refer to a change in meaning, either of a lexical unit (word or expression) or of a concept (a complex knowledge structure that can encompass one or more lexical units, as well as relations among them and with other concepts). The contribution of the proposal to the fields of digital humanities and linguistic data science will therefore consist of a workflow prototype based on a combined approach to semantic change, implying data-related and theoretical enquiry, corpus-based analysis and ontology building, and reflection and documentation on the process as a whole. Since the project is still in an early stage, the paper will limit its scope to the following points: (1) the main elements of the HISTORIAE proposal (goals, tasks, datasets, concepts, challenges); (2) exploratory, preliminary planning and research directions of the intended workflow (theoretical models, formalisms and modalities for detecting and representing semantic change, ontology generation, publication, interpretation and documentation).

1.2 Goals, tasks, methods

HISTORIAE builds on the humanities use case (UC4.2.1) initiated as part of the working group "Use cases and applications" within *Nexus Linguarum, European network for Web-centred linguistic data science*, a COST Action (CA18209) [2] running from 2019 to 2023. While UC4.2.1 will be carried out within *Nexus Linguarum* as a pilot, it is intended to further

[2] `https://nexuslinguarum.eu/`

develop the idea within HISTORIAE as a larger interdisciplinary research project, if funding resources are obtained. The main goal of UC4.2.1 is to create a comparative methodological framework for tracing the "histories" or evolution of concepts in different languages and humanities fields (history, literature, philosophy, religion, etc.) and generate a sample of multilingual LLOD ontologies to represent semantic change by using NLP and Semantic Web technologies. Starting from the hypothesis that historical realities are always reflected in language and its manifestations, irrespective of the specific language, it is assumed that such a methodology will allow for comparative transnational and linguistic standpoints and for new insights into the interconnections between language and historical and cultural context over time and space through linguistic data science.

Six tasks (T1-T6) have been designed for the use case (at the time of writing, T1 is completed, T2, 3, 6 ongoing, T4, 5 not yet started). T1 deals with the identification of potential datasets, concepts and languages to be used in the study. T2 has as objectives to draw on the state-of-the-art in LLOD and NLP methods, tools and data, with a focus on the humanities, and provide a terminological and methodological ground for the construction of a theoretical model to detect and represent semantic change in multilingual historical text collections. T3 consists of the selection of the datasets, periods and time span granularity (years, decades, centuries) as well as data preparation (e.g. conversion from one format to another, grouping by time period). T4 and T5 are dedicated to testing and implementing various methods for semantic change detection, representation and publication as LLOD ontologies based on the selected datasets. Finally, T6 is intended to result interpretation and documentation of the process by making use of explainable AI (XAI) techniques, and to a set of guidelines describing the methodology derived from the use case. More details about the methods considered for further investigation are presented in Section 2.

1.3 Datasets, languages, time span

At the initial stage of the study (T1), we identified several datasets, described below, covering a substantial time span and variety of languages such as Latin, Ancient Greek, Hebrew, French, German, Luxembourgish and Old Lithuanian. The LatinISE corpus [32] contains over 10 million word tokens and covers a wide range of genres (e.g. comedy, tragedy, poetry, essays, letters, narrative, oratory, philosophy, religion, law) spanning from the 2nd century BC to the beginning of the 21st century CE. The corpus is lemmatised, part-of-speech (POS)-tagged and searchable through the Sketch Engine corpus query tool. The Ancient Greek corpus Diorisis [50] covers the Ancient Greek literary tradition, from the 8th century BCE to the 5th century CE, and consists of 820 texts (10,206,421 word tokens), which are lemmatised and POS-tagged. Various genres are represented, such as literature (poetry, drama), philosophy, narrative (historiography, biography, mythography), religion (hymns, Jewish and Christian scriptures, homilies), technical literature (medicine, mathematics, natural science, geography, astronomy, politics, rhetoric, art history, literary criticism, grammar), and letters. The Hebrew dataset Responsa [28] includes rabbinic comments on daily issues (law, health, commerce, marriage, education, Jewish customs) and covers the time range from the 11th century until now. It contains 76,710 articles and about 100 million word tokens and can be browsed and searched via a dedicated Web interface. The National Library of Luxembourg (BnL) Open Data collection [12] comprises historical newspapers and monographs (literature, history, philosophy, geography, pedagogy, religious matters, etc.) from the public domain, in French, German and Luxembourgish. It spans two overlapping periods, 1841-1878 (newspapers) and 1690-1918 (monographs). The dataset counts 23,663 processed newspaper issues (510,505 extracted articles), segmented at the level

of individual articles, sub-articles and paragraphs, and 504 processed monographs (33,477 extracted chapters). The Lithuanian dataset Sliekkas [16], which is still under construction, includes Old Lithuanian texts (religious – prayers, catechisms, hymnals, and sermons, as well as prose and poetry), dated between 16th and 18th centuries, with annotations (structural, paleographic, textological, lexical, and grammatical) and facsimile reproductions of the original (ca. 10 million text words).

1.4 Concepts

Tracing the history of concepts is not a new field of research. Various studies have been dedicated to this area, implying different approaches and domains of application such as political, encyclopaedic, legal and biomedical [51], history of philosophy and of science [4], historical research [13] and digital preservation [47]. However, studies in cultural and conceptual history (Begriffsgeschichte) [1], [41] have pointed out the challenges in examining language in its interaction with social, political and cultural transformations from the real world, and the need for a comparative, transnational and interdisciplinary approach to understand the complexities of this type of relationship.

Our proposal aims at creating comparative standpoints to trace the history of concepts in the domain of socio-cultural transformation at a transnational level. The particularity of the contribution mainly consists in combining various approaches and resources from areas such as the history of concepts and linguistic data science, considered together with this domain of application needing further exploration and insight within a digital framework. A series of semantic fields have been identified for this purpose. Examples of such fields include: geo-political and cultural entities (Europe, West, East, etc.), education, sciences, technology and innovations, social and societal processes (migration, urbanisation, modernisation, globalisation), state and citizenship, beliefs, values and attitudes (e.g. religion, democracy, political participation), economy, health and well-being, everyday life, family and social relations, time and collective memory, work and leisure, customs and traditions, literature and philosophy. Moreover, the study will focus on serendipity and the discovery of "turning points", concepts that underwent significant semantic changes at certain points in time, as indicators of shifting or emerging trends in the area of socio-cultural transformation.

The identified datasets will allow for further research in the history of concepts and within the considered domain. It should be noted that the feature of aligning Old Lithuanian translations with their sources in Latin, German or Polish, or comparing the selected languages, will enable us to identify and assess possible mutual cultural influences and look for emerging shared literary, religious and cultural concepts.

1.5 Challenges

The proposal encompasses a number of challenges. (1) Dataset-related ones mainly referring to aspects such as differences in format, time span, genres, size and availability. (2) Workflow-related ones residing in heterogeneous approaches and workflow components to be integrated into a coherent pipeline; under-developed or not yet existing resources (tools, methods, models, formalisms) to deal with certain languages or aspects of semantic change and diachronic ontology generation and publishing. (3) Domain-related ones generally pointing to questions such as adequacy of the considered datasets and linguistic data science methods and tools for tracing a history of concepts that reflects socio-cultural transformation and provides comparative historical, linguistic and cultural insights from a transnational perspective. Possible modalities of addressing these challenges are described in the following section.

2 Workflow planning

For the implementation of the use case, we propose a workflow composed of seven task categories as illustrated below (Figure 1).

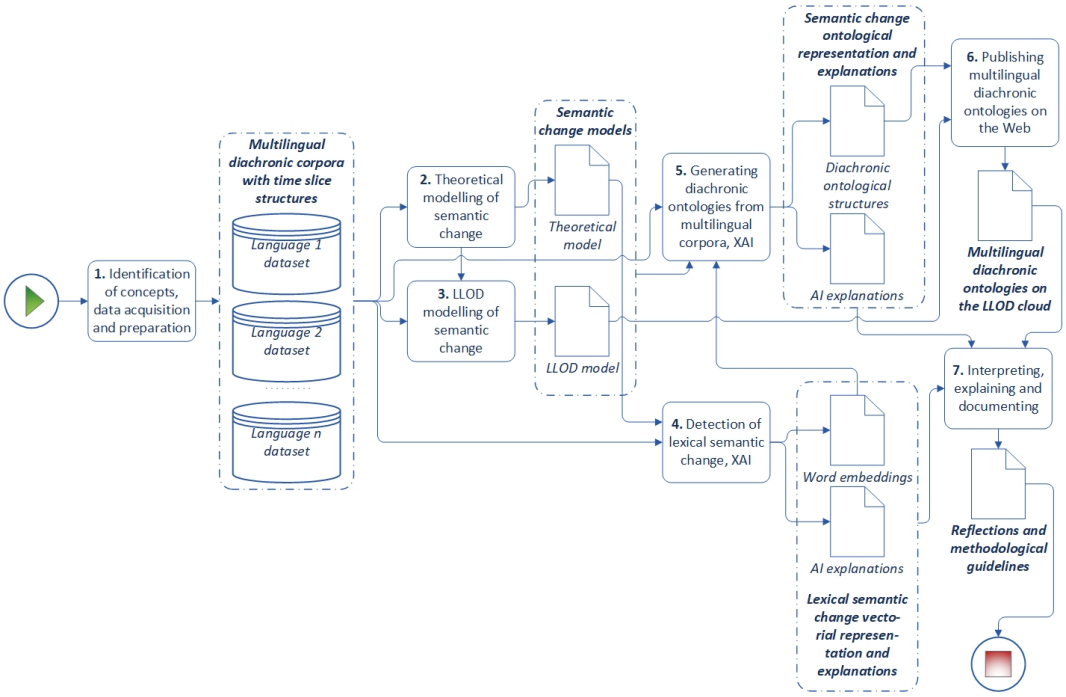

Figure 1 HISTORIAE workflow. Rounded rectangles: task categories; folded-corner rectangles: data types created in the process; dotted rectangles: groups of conceptually associated elements.

2.1 Identification of concepts, dataset acquisition and preparation

An in-depth analysis for the identification of relevant concepts, semantic fields and datasets (T1) (see 1.3, 1.4) to be used in the study will be performed. The core dataset identified so far will be assessed and possibly expanded. The selection criteria mainly pertain to the availability of data, and temporal, geographical, linguistic and thematic coverage enabling a historical, comparative perspective on the topic of socio-cultural transformation. We expect additional datasets, genres and languages to be included in the expanded version of the study (i.e. in Bulgarian, English, Polish, Romanian and Slovene). In order to further extend the time coverage to more recent periods, multilingual contemporary data in open access will be considered, such as Wikimedia downloads, the Digital Corpus of the European Parliament (DCEP) and a collection of trained Twitter word embeddings in English. Preliminary data preparations will be necessary for the whole collection (T3), such as normalisation of old forms, extraction of textual content by genre or language from XML, and segmentation of the corpora by time slice (e.g. year, decade, century) for diachronic analysis.

2.2 Theoretical modelling of semantic change

From a theoretical point of view, four research directions have been identified and will be further explored (T2) as starting points in designing the theoretical model to approach semantic change. It is assumed that such a theoretical model may be combined in the workflow with elements of LLOD formalisation and NLP-based detection of lexical semantic change and diachronic ontology generation (Sections 2.3, 2.4 and 2.5).

Within the theory of lexical semantics, [15] identifies two main classifications of semantic change that include semasiological mechanisms (meaning-related), with semasiological innovations endowing existing words with new meanings, and onomasiological (or "lexicogenetic") (naming-related) mechanisms, with onomasiological innovations expressing meaning through new or alternative lexical items. [15] also draws attention to semantic approaches, in the lineage of distributional semantics, inspired by Firth and Harris, that display a certain affinity with current usage-based approaches and distributional corpus analysis. From the field of intellectual history, theory of knowledge organisation and Semantic Web, two formal descriptions of conceptual change have been retained for further analysis. One is proposed by [27] and asserts that a concept is composed by two parts, the "core" and the "margin", based on context-nonspecific and context-specific features. This model allows for a variety of possibilities, from conceptual continuity, implying core stability and different degrees of margin variability, to conceptual replacement, when the core itself is affected by change. The other formalisation, developed in [51], defines the meaning of a concept in terms of "intension" (a set of properties), "extension" (a subset of the universe) and "label" (a string of characters). [51] use distance measures, such as Jaccard and Levenshtein, computed for the three aspects to identify conceptual changes.

2.3 Expressing semantic change through LLOD formalisms

One of the aims of the work described in this submission is to model and then publish data about semantic change in the form of one or more diachronic lexico-ontologies in order to integrate together different kinds of relevant information and to make this information available in an accessible and easily re-usable form. The linked open data (LOD) publishing paradigm is ideal for doing this. It offers us a standardised way of making structured data available using the HTTP protocol, as well as giving us the possibility of exposing this data via special endpoints that use the powerful SPARQL query language. The use of a common data framework, the Resource Description Framework (RDF), combined with a number of upper level ontologies and more generalised linked data vocabularies helps to ensure the interoperabilty of data published in this way. As we intend to model (and publish) data about linguistic phenomena as linked data (although this may include information from and relevant to other disciplines such as history) we use the term linguistic linked open data in the current work. In the rest of this section we will give a brief overview (T2) of some of the most relevant vocabularies and datasets for publishing data on semantic change as linguistic linked open data.

The idea would be to create a linked data resource with a lexical component that includes a list of lexical entries and their senses (along with other linguistic information pertaining to for instance the grammatical features of a word) and an ontological or more broadly speaking semantic component that describes the meanings of these senses and, more importantly, the way in which they change over time. The well known OntoLex-Lemon model [31] published by the W3C Ontology-Lexicon group[3] allows for this approach in the case of static senses. However it does not make explicit provision for representing semantic change, nor does it do so for dynamic or time dependent information. This is an issue because the representation of n-ary relations for $n > 2$ can have its drawbacks [52] (in this case relationships which would be most naturally represented as relations with an additional temporal parameter).

[3] `https://www.w3.org/2016/05/ontolex/`

The modelling of dynamic or diachronic lexical information in linked data is still an active area of research and discussion, and it is unlikely that there will be any one-size-fits-all solution. One approach has been proposed in [23] where word senses are represented as perdurants, that is, entities with an extension in time which can have temporal parts. This strategy is also being adopted in the soon to be published ISO Standard ISO 24613-3 which consists of a diachronic module for the Lexical Markup Framework (LMF) [43]. In the case of LLOD the perdurant solution has the advantage, among other things, of allowing the use of certain built-in Web Ontology Language constructs which facilitate automated reasoning on RDF datasets.

2.4 Detecting lexical semantic change

There is a growing body of research on computational methods for detecting lexical semantic change automatically, recently surveyed in [48] and [26].

Word representations that employ semantics, syntax, and context to create vectors are used in current literature to successfully compute semantic change using distance metrics (e.g., cosine, Levenshtein) [46]. These vectors are built using shallow neural networks, and, although they use different architectures to create lexical representation for textual data, are known collectively as word embeddings [34, 38, 37, 6, 35]. Although similar concepts have similar representations, word embeddings cannot detect correctly the semantic changes that appear over time if they are not trained specifically for this task. Thus, in current literature new methods for building word embeddings to detect semantic changes have been proposed. In [18], the authors correlate word embeddings with temporal-spatial information to create condition-specific embeddings. Another method uses hyperbolic embeddings to map partial graphs into low-dimensional, continuous hierarchical spaces to build diachronic semantic hyperspaces for four scientific topics [5]. The current approaches are prone to anomalies and direct human intervention is required to make correct assessments about the results. Thus, new anomaly detection methods that employ unsupervised machine and deep learning are required to alleviate the need for expert validation.

Word embeddings are used with machine translation architectures, e.g., long short-term memory networks (LSTM)-based sequence to sequence models [49], to measure the semantic change of words by tracking their evolution over time in a sequential manner. These approaches seem promising and new deep learning architectures for machine translation can be developed for the task of determining semantic change, e.g., variational autoencoders (VAE) [24] and generative adversarial networks (GAN) [29].

The SemEval 2020 shared task on Unsupervised Lexical Semantic Change Detection [46] has provided the current state-of-the-art in the field, with evaluation results relative to 21 systems evaluated on two subtasks in four languages (English, German, Latin and Swedish). In this task, systems based on word type embeddings outperformed token embeddings on both subtasks, but the potential of token embeddings is yet to be fully explored. [46] also found a strong effect of frequency in the systems based on type embeddings, and a strong correlation between change scores and polysemy. Both these factors should be further explored and taken into account in future studies and implementations.

Recently transformers-based models have also been considered for lexical semantic change detection. Most solutions that fall into this category use BERT (Bidirectional Encoder Representations from Transformers) [10]. BERT employs a bidirectional attention mechanism to learn the contextual relations. Pre-trained BERT models have been used in both unsupervised [17, 22] and supervised [42] semantic shift tracing solutions. Another transformer that was applied to semantic change detection is ELMo (Embeddings from Language

Models) [39]. ELMo provides faster training and inference compared with BERT. Because of this, it is much easier to train the models with ELMo on specific datasets, and not use pre-trained models. A comparison between ELMo and BERT in semantic change detection for the Russian language is presented in [42]. By analysing the results presented in the state-of-the-art solutions, we can conclude that transformers enhance the semantic change detection task. For this purpose, we are considering experimenting with other, more recent transformers (T4). One candidate is DistilBERT [45] which is used to pre-train a smaller general-purpose language representation model by reducing the size of BERT. RoBERTa [30] is another candidate. This model improves BERT's language masking strategy by adjusting several hyperparameters.

2.5 Generating diachronic ontologies from corpora

Another area of interest for the study is that of ontology learning from text, surveyed in [21, 2]. An influential model used in many applications, the so-called "ontology learning layer cake" [7], proposes six steps or layers for ontology acquisition, dedicated to *terms*, *synonyms*, *concepts*, *concept hierarchies*, *relations* and *rules*. [7] list different techniques from various fields of research to achieve this task. The first three subtasks include information retrieval methods for term extraction, synonym acquisition from lexical-semantic resources (e.g. WordNet), text corpora and the Web based on synset relations, Harris' distributional hypothesis and statistical information measures, and concept induction through definition learning (*intension*), deriving instances from named entity hierarchies (*extension*) and linguistic realisation (*terms*) (see also Section 2.2). For the last three subtasks, [7] mention taxonomic (*is-a*) and non-taxonomic relation extraction based on hierarchical clustering algorithms as well as statistical and linguistic analysis of syntactic structure and dependencies, and ontological rules learning from text using lexical entailment. This framework will serve as a starting point for designing this workflow phase (T5), possibly in combination with deep learning approaches (e.g. word2vec, LSTMs), human-based evaluation and post-processing that seem promising for ontology learning goals according to [53], [21] and [2].

For more specific objectives in generating diachronic ontologies from historical corpora included in the use case, additional methods will be assessed, such as distributional semantic models and "hubs and authorities" [20], hyperbolic embeddings [5], "peak detection" in time series and word and event "projected embeddings" [44] or vector representation of concept signatures [19]. Possible integration with recent advances in transformers-based models and other state-of-the-art NLP methods for lexical semantic change detection (Section 2.4) will be considered as well.

2.6 Publishing diachronic ontologies as LL(O)D

Unlike synchronic ontologies that ignore the historical perspective, diachronic ontologies allow us to capture the temporal dimension of concepts and investigate gradual semantic changes and concept evolution through time [20]. Since the goal of the study is to produce a sample of diachronic ontologies represented and published on the Web as LL(O)D (T5), a set of existing methods and tools for acquiring (Section 2.5) and converting ontological structures into Semantic Web formalisations will be evaluated together with modalities of expressing semantic change through LLOD formalisms (Section 2.3). One of the systems often cited as a reference is Text2Onto [9], an ontology learning framework that converts learned knowledge into a Probabilistic Ontology Model (POM) translatable into various ontology representation languages such as RDFS, OWL and F-Logic. Other tools, e.g. LODifier [3] and OntoGain [11],

can extract entities and relations from text and produce RDF representations linked to the LOD cloud using DBpedia and WordNet 3.0 vocabularies, or transform the acquired ontology into standard OWL statements. More specialised tools, such as converters, allow for making linked data in RDF format out of CSV files (CoW [33]), converting language resources into LLOD (LLODifier [8]) or developing complex transformation pipelines for converting heterogeneous linguistic resources to RDF (Fintan [14]).

2.7 Interpreting, explaining and documenting the process

For the interpretative approach, we will take into account linguistic, cultural and historical aspects of linguistic innovation and its temporal and referential complexity starting from the theoretical model of the *concept – reality relationship*, based on the four combinations of synchronous and asynchronous concept and reality change vs. stability over a period of time [25]. The implementation of the proposed workflow will include qualitative analysis and XAI components (in phases 2.4 and 2.5) for interpreting the results, explaining, documenting and reflecting on the process (T6). As starting points we will consider the four principles of explainable AI systems [40] and insight from the social sciences in designing this type of components [36]. The outcome will consist of comparative insights into the history of socio-cultural transformation, and in particular the interconnection between linguistic innovation and social and cultural innovation, and their evolution over time. It will also contain methodological guidelines and reflections on the hybridisation of human and algorithmic approaches and the role of AI from the sociology of knowledge perspective, in order to understand how these technologies are changing our modes of producing, disseminating and consuming knowledge.

3 Conclusion and future work

The paper presents a use case and further development proposal for detecting and representing semantic change by means of NLP and LL(O)D technologies applied to multilingual historical datasets and various humanities areas in order to trace the evolution of concepts in the domain of socio-cultural transformation. A set of challenges has been identified, mainly related to the heterogeneity of the datasets and approaches, as well as the complexity of the application domain and of constructing comparative standpoints to derive historical, linguistic and cultural insight from a transnational perspective. Given the early stage in the use case development, the proposal does not present experimental descriptions and results but a set of methodologies and tools to be further examined, tested and evaluated within the planned workflow. It is expected that some of the defined challenges will be addressed by combining various approaches in linguistic data science, e.g. for theoretical modelling, detection and representation of semantic change and diachronic ontology learning, as well as documentation and reflection on the process itself making use of human- and AI-based explainability. The dataset diversity may provide opportunities for reflection on the gaps in the data and the possibilities for alleviating incompleteness and uncertainty by a modular, expansible design and an explainability- and discovery-based architecture. The next steps of the study will therefore consist in testing the hypotheses formulated in the present proposal to confirm or disconfirm their validity and create the bases for the construction of the comparative framework and workflow prototype for detecting and representing semantic change through NLP and LLOD technologies.

References

1 Alessandro Arcangeli. *Cultural History. A Concise Introduction.* Routledge, 1 edition, 2012. doi:10.4324/9780203789247.

2 Muhammad Nabeel Asim, Muhammad Wasim, Muhammad Usman Ghani Khan, Waqar Mahmood, and Hafiza Mahnoor Abbasi. A Survey of Ontology Learning Techniques and Applications. *Database*, 2018, January 2018. doi:10.1093/database/bay101.

3 Isabelle Augenstein, Sebastian Padó, and Sebastian Rudolph. LODifier: Generating Linked Data from Unstructured Text. In Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti, editors, *The Semantic Web: Research and Applications*, volume 7295 of *Lecture Notes in Computer Science*, page 210–224. Springer Berlin Heidelberg, 2012. doi:10.1007/978-3-642-30284-8_21.

4 Arianna Betti and Hein van den Berg. Modelling the History of Ideas. *British Journal for the History of Philosophy*, 22(4):812–835, 2014. doi:10.1080/09608788.2014.949217.

5 Yuri Bizzoni, Marius Mosbach, Dietrich Klakow, and Stefania Degaetano-Ortlieb. Some Steps Towards the Generation of Diachronic WordNets. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 55–64, 2019. URL: https://www.aclweb.org/anthology/W19-6106.

6 Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi:10.1162/tacl_a_00051.

7 P. Buitelaar, P. Cimiano, and B. Magnini. Ontology Learning from Text: An Overview. In *Ontology Learning from Text: Methods, Evaluation and Applications*, volume 123, pages 3–12. IOS Press, 2005.

8 Philipp Cimiano, Christian Chiarcos, John P. McCrae, and Jorge Gracia. Linguistic Linked Data in Digital Humanities. In *Linguistic Linked Data. Representation, Generation and Applications*, pages 229–262. Springer International Publishing, 1 edition, 2020. URL: https://www.springer.com/gp/book/9783030302245.

9 Philipp Cimiano and Johanna Volker. 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*, 2005.

10 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019. doi:10.18653/v1/N19-1423.

11 Euthymios Drymonas, Kalliopi Zervanou, and Euripides G. M. Petrakis. Unsupervised Ontology Acquisition from Plain Texts: The OntoGain System. In Christina J. Hopfe, Yacine Rezgui, Elisabeth Métais, Alun Preece, and Haijiang Li, editors, *Natural Language Processing and Information Systems*, volume 6177 of *Lecture Notes in Computer Science*, page 277–287. Springer Berlin Heidelberg, 2010. doi:10.1007/978-3-642-13881-2_29.

12 Maud Ehrmann, Matteo Romanello, Simon Clematide, Phillip Ströbel, and Raphaël Barman. Language Resources for Historical Newspapers: the Impresso Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA), 2020.

13 Antske Fokkens, Serge Ter Braake, Isa Maks, and Davide Ceolin. On the Semantics of Concept Drift: Towards Formal Definitions of Semantic Change. *Drift-a-LOD@EKAW*, 2016.

14 Christian Fäth, Christian Chiarcos, Björn Ebbrecht, and Maxim Ionov. 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.

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

16 Jolanta Gelumbeckaite, Mindaugas Sinkunas, and Vytautas Zinkevicius. Old Lithuanian Reference Corpus (SLIEKKAS) and Automated Grammatical Annotation. *J. Lang. Technol. Comput. Linguistics*, 27(2):83–96, 2012.

17 Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 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, 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.365.

18 Hongyu Gong, Suma Bhat, and Pramod Viswanath. Enriching Word Embeddings with Temporal and Spatial Information. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 1–11, Online, November 2020. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/2020.conll-1.1.

19 Jon Atle Gulla, Geir Solskinnsbakk, Per Myrseth, Veronika Haderlein, and Olga Cerrato. Semantic Drift in Ontologies. In *WEBIST 2010, Proceedings of the 6th International Conference on Web Information Systems and Technologies*, volume 2, April 2010.

20 Shaoda He, Xiaojun Zou, Liumingjing Xiao, and Junfeng Hu. Construction of Diachronic Ontologies from People's Daily of Fifty Years. *LREC 2014 Proceedings*, 2014.

21 Vivek Iyer, Mohan Mohan, Y. Raghu Babu Reddy, and Mehar Bhatia. A Survey on Ontology Enrichment from Text. In *The sixteenth International Conference on Natural Language Processing (ICON-2019)*, 2019.

22 Vani Kanjirangat, Sandra Mitrovic, Alessandro Antonucci, and Fabio Rinaldi. SST-BERT at SemEval-2020 Task 1: Semantic Shift Tracing by Clustering in BERT-based Embedding Spaces. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 214–221, 2020.

23 Anas Fahad Khan. Towards the Representation of Etymological Data on the Semantic Web. *Information*, 9(12), November 2018. Publisher: MDPI AG. doi:10.3390/info9120304.

24 Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014.

25 Reinhart Koselleck. Some Reflections on the Temporal Structure of Conceptual Change. In Willem Melching and Velema Wyger, editors, *Main Trends in Cultural History. Ten Essays*, page 7–16. Rodopi, 1994.

26 Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic Word Embeddings and Semantic Shifts: A Survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics.

27 Jouni-Matti Kuukkanen. Making Sense of Conceptual Change. *History and Theory*, 47(3):351–372, 2008. doi:10.1111/j.1468-2303.2008.00459.x.

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

29 Jianyi Liu, Yu Tian Ru Zhang, Youqiang Sun, and Chan Wang. A Two-Stage Generative Adversarial Networks With Semantic Content Constraints for Adversarial Example Generation. *IEEE Access*, 8:205766–205777, 2020. doi:10.1109/ACCESS.2020.3037329.

30 Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*, July 2019. arXiv:1907.11692.

31 John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. The OntoLex-Lemon Model: Development and Applications. *Electronic Lexicography in the 21st Century. Proc. of eLex 2017 conference, in Leiden, Netherlands*, pages 587–597, September 2017. Publisher: Lexical Computing CZ s.r.o. URL: https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf.

32 Barbara McGillivray and Adam Kilgarriff. Tools for Historical Corpus Research, and a Corpus of Latin. *New Methods in Historical Corpus Linguistics*, 1(3):247–257, 2013.

33 Albert Meroño-Peñuela, Victor de Boer, Marieke van Erp, Willem Melder, Rick Mourits, Ruben Schalk, and Richard Zijdeman. Ontologies in CLARIAH: Towards Interoperability in History, Language and Media. *arXiv*, 2020. arXiv:2004.02845v2.

34 Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations*, pages 1–12, 2013.

35 Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in Pre-Training Distributed Word Representations. In *International Conference on Language Resources and Evaluation*, pages 52–55, 2018.

36 Tim Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267:1–38, February 2019. doi:10.1016/j.artint.2018.07.007.

37 Maximilian Nickel and Douwe Kiela. Poincaré Embeddings for Learning Hierarchical Representations. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6341–6350, 2017.

38 Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, October 2014. doi:10.3115/v1/D14-1162.

39 Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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, 2018. Association for Computational Linguistics. doi:10.18653/v1/N18-1202.

40 P. Jonathon Phillips, Carina A. Hahn, Peter C. Fontana, David A. Broniatowski, and Mark A. Przybocki. *Four Principles of Explainable Artificial Intelligence*. National Institute of Standards and Technology, U.S. Department of Commerce, August 2020. doi:10.6028/NIST.IR.8312-draft.

41 Melvin Richter. *The History of Political and Social Concepts: A Critical Introduction*. Oxford University Press, 1995.

42 Julia Rodina, Yuliya Trofimova, Andrey Kutuzov, and Ekaterina Artemova. ELMo and BERT in Semantic Change Detection for Russian. *CoRR*, abs/2010.03481, 2020. arXiv:2010.03481.

43 Laurent Romary, Mohamed Khemakhem, Fahad Khan, Jack Bowers, Nicoletta Calzolari, Monte George, Mandy Pet, and Piotr Bański. LMF Reloaded. *arXiv preprint*, 2019. arXiv:1906.02136.

44 Guy D. Rosin and Kira Radinsky. Generating Timelines by Modeling Semantic Change. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, page 186–195. Association for Computational Linguistics, 2019. doi:10.18653/v1/K19-1018.

45 Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. In *Workshop on Energy Efficient Machine Learning and Cognitive Computing*, pages 1–5, 2019.

46 Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, 2020.

47 Thanos G Stavropoulos, Stelios Andreadis, Marina Riga, Efstratios Kontopoulos, Panagiotis Mitzias, and Ioannis Kompatsiaris. A Framework for Measuring Semantic Drift in Ontologies. In *CEUR Workshop Proceedings Vol-1695*, September 2016.

48 Nina Tahmasebi, L. Borin, and A. Jatowt. Survey of Computational Approaches to Lexical Semantic Change. *arXiv*, 2018. arXiv:1811.06278.

49 Adam Tsakalidis and Maria Liakata. 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, November 2020. doi:10.18653/v1/2020.emnlp-main.682.

50 Alessandro Vatri and Barbara McGillivray. The Diorisis Ancient Greek Corpus: Linguistics and Literature. *Research Data Journal for the Humanities and Social Sciences*, 3(1):55–65, 2018.

51 Shenghui Wang, Stefan Schlobach, and Michel Klein. Concept Drift and How to Identify It. *Journal of Web Semantics First Look*, September 2011. doi:10.2139/ssrn.3199520.

52 Chris Welty, Richard Fikes, and Selene Makarios. A Reusable Ontology for Fluents in OWL. In *FOIS*, volume 150, pages 226–236, 2006.

53 Gerhard Wohlgenannt and Filip Minic. Using word2vec to Build a Simple Ontology Learning System. *International Semantic Web Conference*, page 4, 2016.