UDK 81'322
81'37
Izvorni znanstveni rad
Rukopis primljen 26. IV. 2023.
Prihvaćen za tisak 16. XII. 2024.
doi.org/10.31724/rihjj.50.2.1

Florentina Armaselu[1]
Barbara McGillivray[2]
Chaya Liebeskind[3]
Paola Marongiu[4]
Giedrė Valūnaitė Oleškevičienė[5]
Elena-Simona Apostol[6]
Ciprian-Octavian Truică[7]
florentina.armaselu@uni.lu
barbara.mcgillivray@kcl.ac.uk
liebchaya@gmail.com
paola.marongiu@ilc.cnr.it
gvalunaite@mruni.eu
elena.apostol@upb.ro
ciprian.truica@upb.ro

MULTILINGUAL WORD EMBEDDING AND LINGUISTIC LINKED OPEN DATA FOR TRACING SEMANTIC CHANGE

This article proposes a methodology for combining natural language processing techniques for diachronic analysis and linguistic linked open data models to detect and represent semantic change. The change in meaning over time of words, phrases, or concepts encompasses complex phenomena that cannot be fully ex-

[1] University of Luxembourg, Luxembourg; [2]King's College London, United Kingdom; [3]Jerusalem College of Technology, Israel; [4]Institute of Computational Linguistics "Antonio Zampolli", National Research Council (ILC-CNR), Italy; [5]Mykolas Romeris University, Lithuania; [6]National University of Science and Technology Politehnica Bucharest, Romania; [7]National University of Science and Technology Politehnica Bucharest, Romania.
Corresponding author Florentina Armaselu, orcid.org/0000-0003-2386-6889.

plained by distributional methods alone. We argue that by joining corpus-based and lexicographical evidence and modelling the results in an interoperable format can provide more solid ground for drawing conclusions and possibilities of re-use by other applications. We define a basic schema for a resource aggregator and a model called LLODIA (Linguistic Linked Open Data for Diachronic Analysis). To illustrate our approach, we use a multilingual dataset, in French, Latin, Hebrew, Old Lithuanian, and Romanian, and build a sample derived from word embeddings and dictionary resources, encoded by means of the proposed model.

1. Introduction

The aim of the present article is to illustrate how corpus- and dictionary-based techniques can be brought together with linked data models to detect and represent semantic change. We use the term *semantic change* in a broad sense, referring to a change in meaning over time of a lexical unit (word or expression) or concept (a complex knowledge structure that can encompass one or more lexical units and relations among them and with other concepts). The study of semantic change is an open research challenge in historical linguistics and the digital humanities (DH) and involves addressing questions such as: What are the mechanisms that determine semantic innovation? Do they involve sudden shifts or slow processes? How are they related to the real world? How can the gradual change in word meaning and its attestation at discrete points in time be captured? Is it possible to trace and model this type of phenomenon through natural language processing (NLP) methods and linguistic linked open data (LLOD)[2] formalisms, and what types of resources are needed for this purpose? While the first queries are indirectly addressed as a general framework, in this study, we focus on the last two questions. We propose to combine word embedding techniques, dictionary evidence and LLOD modelling (Khan et al. 2022) to trace the evolution of a set of concepts in a collection of multilingual diachronic corpora.

[2] A set of principles, resources and a community dealing with the creation and publication on the Web of open data for linguistics and natural language processing. See: https://linguistic-lod.org/.

Our outcomes are an open-source LLOD model and a sample conceived as a proof of concept of our approach.

1.1. State of the art

In his analysis[3] of the temporal structure of conceptual change, Koselleck (1994: pp. 9, 14–16) reflects on the relationship between concepts and reality, which involves four states of stability and change (concepts and reality changing or remaining stable together, one changing and the other remaining stable, and vice-versa). He distinguishes various types of source materials that can be used to reconstruct the history of these changes. One type refers to sources such as newspapers, letters, memoranda, and speeches, dedicated to instant consumption and possessing a single temporal layer. The other type includes normative sources such as lexicons, dictionaries, encyclopaedias, and handbooks that display a slow evolution and represent indispensable tools in tracing the gradual development of new layers of meaning and the pace of conceptual change in a certain language and culture. Similarly, Richter (1994: p. 125) insists on the importance and diversity of the resources to be used in tracing the history of concepts and on the interaction between conceptual and social changes as a reflection of reality, when language is understood both as an agent and an indicator of change. Therefore, this type of research should use a broad range of materials, such as newspapers, journals, official documents, memoirs, correspondence, diaries, and literary texts, but also encyclopaedias, lexicons, handbooks, and thesauri, in order to capture the ways in which language shapes and registers the process of change. Particular attention is paid by Richter to the alternation between diachronic and synchronic analyses of language, and the use of both semasiological and onomasiological standpoints for the study of all the meanings of a term or concept, and respectively of all the names or terms in a language corresponding to the same thing or concept.

While these studies offer theoretical insights into the types of sources and methodological aspects to be considered, from the computational processing perspective, corpus analysis, based on the principles of distributional semantics, and the

[3] For a general discussion on the history of concepts, see for instance (Kuukkanen 2008, Gavin et al. 2019).

use of word embeddings for detecting lexical semantic change have shown advances in recent years (Tahmasebi et al. 2021, Schlechtweg et al. 2020, Kutuzov et al. 2018, Chiru et al. 2021). Research has also been carried out to address some of the drawbacks of the embedding techniques, e.g., the meaning conflation deficiency and related interpretability issues, through more complex, contextual models, such as ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019), or by combining unsupervised and knowledge-based methods for computing sense embeddings and fine-grained representations of meaning (Camacho-Collados and Pilehvar 2018, Hu et al. 2019, Scarlini et al. 2020, McGillivray et al. 2023) that may be used in diachronic analysis.

LLOD studies in modelling etymologies (Declerck et al. 2015, Abromeit et al. 2016, Khan 2020) and proposals for linking lexical resources with digital editions and corpus evidence (Tittel et al. 2018, Chiarcos et al. 2022a, Chiarcos et al. 2023) have also shown recent progress. However, there are currently not many initiatives that bridge the NLP and LLOD domains for tracing and modelling semantic change (Armaselu et al. 2022). There is also no general agreement on how to apply Semantic Web formalisms to model semantic change. Previous research focused, for instance, on the theoretical aspects of representing concept drift in ontological resources such as DBpedia, by tracing changes in concept intension, extension and label via RDF mechanisms and similarity/distance measures (Wang et al. 2011, Fokkens et al. 2016). Other researchers addressed the question by means of RDF-based etymologies and temporally-enriched dictionaries (Khan 2020). Basile et al. (2022) applied the GraphBRAIN Schema graph database format to model relations between concepts and words, information about word occurrences, and diachronic information about concepts and words.

1.2. Contribution

The theoretical background and methodological insights provided by Koselleck and Richter can serve as guidelines for imagining more inclusive approaches in analysing semantic change, in combination with NLP and LLOD techniques. Our proposal will, therefore, investigate different ways of combining various resources and methodologies in order to capture (even if partially) the complexity of the studied phenomena as characterised by their linguistic and factual, reality-based aspects. For this, both “normative” or “descriptive” (e.g., dictionaries) and language-in-context (e.g., corpora) sources should be exploited. Furthermore, we argue that the coverage, representation, and interconnection capability of the Semantic Web and the linked data paradigm are flexible and powerful enough to support the development of a framework allowing for modelling, querying, and reasoning related to the representation of word meaning evolution over time. This contribution focuses on the creation of a semantic model and multilingual proof of concept for tracing semantic change by aggregating various resources and methods. Section 2 illustrates our experiments with the core dataset. Sections 3 and 4 discuss the proposed LLOD model and outline a possible resource aggregator inspired by it, and section 5 summarises our findings and future work. Figure 1 shows the main phases of the project pipeline.

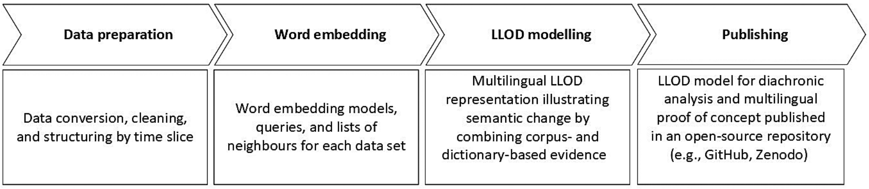

Data preparation
Word embedding
LLOD modelling
Publishing
Data conversion, cleaning, and structuring by time slice
Word embedding models, queries, and lists of neighbours for each data set
Multilingual LLOD representation illustrating semantic change by combining corpus- and dictionary-based evidence
LLOD model for diachronic analysis and multilingual proof of concept published in an open-source repository (e.g., GitHub, Zenodo)

Figure 1: Project pipeline. Main phases and results

2. Methodology

The core dataset we used (Figure 2) contains various types of time span, genres, and documents in five extinct and extant languages (Latin, Hebrew, Old Lithuanian, French, and Romanian). The methodology consists of the following steps: (1) preprocessing the selected corpora (data/metadata extraction, conversion, structuring by time slice); (2) training word embeddings by time slice for each language; (3) querying the resulting models for terms or groups of terms, and extracting similarities and examples of context by language and time period; (4) comparing the results across the languages in the dataset, and with multilingual historical dictionaries, when available; (5) creating a set of interrelated multilingual LLOD representations, combining corpus evidence and lexicographical sources to trace semantic change. Some of these steps are standard in processing multilingual linked open data (Gromann et al. 2024) using big data approaches (Trajanov et al. 2024).

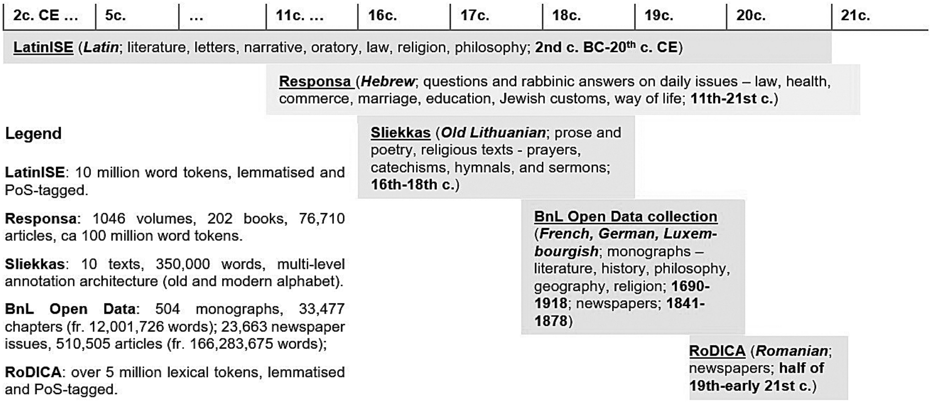

2c. CE …
5c.
…
11c. …
16c.
17c.
18c.
19c.
20c.
21c.
LatinISE (Latin; literature, letters, narrative, oratory, law, religion, philosophy; 2nd c. BC-20th c. CE)
Responsa (Hebrew; questions and rabbinic answers on daily issues – law, health, commerce, marriage, education, Jewish customs, way of life; 11th-21st c.)
Sliekkas (Old Lithuanian; prose and poetry, religious texts - prayers, catechisms, hymnals, and sermons; 16th-18th c.)
BnL Open Data collection (French, German, Luxembourgish; monographs – literature, history, philosophy, geography, religion; 1690-1918; newspapers; 1841-1878)
RoDICA (Romanian; newspapers; half of 19th-early 21st c.)
Legend
LatinISE: 10 million word tokens, lemmatised and PoS-tagged.
Responsa: 1046 volumes, 202 books, 76,710 articles, ca 100 million word tokens.
Sliekkas: 10 texts, 350,000 words, multi-level annotation architecture (old and modern alphabet).
BnL Open Data: 504 monographs, 33,477 chapters (fr. 12,001,726 words); 23,663 newspaper issues, 510,505 articles (fr. 166,283,675 words);
RoDICA: over 5 million lexical tokens, lemmatised and PoS-tagged.

Figure 2: Core dataset: time span, language, genre and size of its components.

Our goal is to trace and compare the evolution of a number of (parallel) concepts reflecting sociocultural and historical aspects by using a multilingual diachronic collection, and to get insights into the process of linguistic innovation and change from a cross-language, cross-cultural perspective. The corpora were selected based on criteria such as linguistic variety (including under-represented languages), a long enough time span to allow for a diachronic perspective, and availability of the texts either in open access or through specific research agreements. For the experiments described in this paper we applied embedding techniques, such as word2vec, for French and Hebrew, fastText, for Latin and Old Lithuanian, the latter reported to work better for these two languages than the former, and word2vec and ELMo for Romanian. It is also intended to show how the results of word embeddings using multilingual diachronic corpora can be combined with attestation, lexico-semantic and etymological information from dictionaries and be modelled via LLOD formalisms such as OntoLex-FrAC[4] and the proposed model LLODIA.

[4] An extension of the OntoLex Lexicon Model for Ontologies (OntoLex-Lemon), intended to enrich lexical resources with corpus-based information such as attestations via corpus examples, frequency and collocation scores, embeddings, and similarity metrics. See: https://www.w3.org/community/ontolex/wiki/Frequency,_Attestation_and_Corpus_Information.

2.1. BnL Open Data

The BnL Open Data[5] corpus (Ehrmann et al. 2020) (Figure 2) contained the TEXT ANALYSIS PACK and the MONOGRAPH TEXT PACK, two sets of downloadable historical data (newspapers and monographs in the public domain and in several languages) provided by the Bibliothèque nationale du Luxembourg (BnL) (National Library of Luxembourg). The data was in XML following the Dublin Core format (processed via the BnLMetsExporter[6]) and contained metadata (persistent ARK[7] identifier, source, title, publisher, date, language, etc.) and text content for units such as newspaper articles, book chapters, and advertisements. In the preprocessing phase, the text was extracted from the initial hierarchy of folders and the XML files, and stored in a plain text format, with the names of the new files including a prefix indicating the language and date of publication of the text. This type of information was subsequently used to restructure the dataset by time slice, according to the date of certain events considered important for the history of Luxembourg and that potentially could have determined linguistic changes at the time (e.g., the invasion of the Napoleonic troops, independence, the dismantling of the fortress of Luxembourg, withdrawal of the Prussian garrison or royal decrees and school laws stating the official languages of the Grand Duchy and the languages taught in school). The preparation of the text versions (with metadata extracted in CSV files), the selection of the French documents for analysis, and the structuring of the dataset by time slice was performed automatically using the workflow management platform KNIME [8].

Word2vec (Mikolov et al. 2013, Řehůřek and Sojka 2010) was applied by time slice on the BnL dataset (5 word context window, 100 dimension vectors). Table 1 shows the structure of the BnL monograph dataset (that was used in this study) and the size of the vocabulary for the word embedding models generated for each time interval. One can notice the small vocabulary of the models, especially for the period 1690-1830, which signals document scarcity for these intervals. OCR problems (confusion of 'f'/'s', 'c'/ 'e', 'n'/'u', 'l'/'t'), incorrect assignation of docu-

[5] https://data.bnl.lu/data/historical-newspapers/.

[6] https://data.bnl.lu/tools/#bnlmetsexporter.

[7] https://arks.org/

[8] https://www.knime.com/.

ment language, issues with the tokenisation, hyphenation, lemmatisation, and French stopwords were also observed. In some cases, incorrect words tended to have higher similarity with the corresponding correct ones, as already reported by (Van Strien et al. 2020).

Table 1: BnL. French monographs by time slice and vocabulary size (lemmas) of the models

Word embedding model	Vocabulary size
1690 - 1794	1,924
1795 - 1814	318
1815 - 1830	2,211
1831 - 1866	32,842
1867 - 1889	46,549
1890 - 1918	21,105

The most similar lemmas (neighbours) were computed via cosine similarity[9] for a set of targets selected empirically from semantic fields related to various themes such as socio-political, cultural, institutional, military, economic, administrative and language-related domains.

Once the corresponding neighbours were obtained for the target words, the question was if the differences observed in the neighbour sets for each target and time slice were reflecting semantic change, at the time when it was emerging, or different cases of polysemy already attested by other sources. For comparison, we looked first at the French Wiktionary[10], given its broad coverage of languages and types of information provided for the entries (e.g., etymology, part of speech, definitions and examples for different senses, translations).

Table 2 presents the example for the lemma *grain*[11], its neighbours, two excerpts from the corpus as digital facsimiles, and the corresponding definitions in Wiktionary for French and English. The neighbours point to the senses *unit of weight* and *seed*. Although Wiktionary is a rich resource, which often contains citations with indications of the sources, it didn't allow for a systematic comparison be-

[9] We also computed a pair-based similarity measure to bring to the top of the list neighbours with higher mutual similarity.

[10] https://fr.wiktionary.org.

[11] *Grain*: lemma frequency for the whole corpus: 902; time slice frequency: 1815-1830: 9; 1890-1918: 158.

tween the date of the occurrence in the corpus and an attestation date.Table 2: BnL French monographs: *grain*, top neighbours by time slice, corpus context and Wiktionary entries

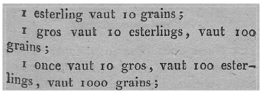
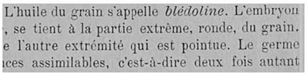
grain (1815-1830)	'gros', 'verre', 'mesurettes', 'boisseau', 'lest', 'hectolitre', 'chopine', ...	Tables de conversion des poids,[11] J.Lamort, Luxembourg, 1820 (p. 23)[12]  1 esterling vaut 10 grains ; 1 gros vaut 10 esterlings, vaut 100 grains ; 1 once vaut 10 gros, vaut 100 esterlings, vaut 1000 grains ;	Wiktionary.grain fr.[13] *(Métrologie)* *(Vieilli) Poids qui était le soixante-douzième d'un gros.* en.[14] 8. *Any of various small units of length originally notionally based on a grain's width ...*
grain (1890-1918)	'maltine', 'méteil', 'blédoline', 'battage', 'épeautre', 'féculent', 'tarare', ...	L'Abbé N. Neuens, L'hygiène de la table,[15]A. Woitrin, Namur, 1898 (p. 13)[16]  L'huile du grain s'appelle *blédoline*. L'embryon , se tient à la partie extrême, ronde, du grain. e l'autre extrémité qui est pointue. Le germe ces assimilables, c'est-à-dire deux fois autant	Wiktionary.grain fr.[17] 1. *Fruit et semence des céréales contenu dans l'épi.* en.[18] 1. *The harvested seeds of various grass food crops ...*

Table 3 shows another example of neighbour lists and corpus excerpts combined with the CNRTL[20]-Ortolang lexical portal entries that include attestation dates. The senses from the dictionary were semi-automatically assigned, using ChatGPT-4 (OpenAI 2023), to match the meaning conveyed by sub-sets of neighbours resulting from the word embedding analysis and examples of contexts manually selected from the corpus (Armaselu et al. 2024c).[21]

[12] https://viewer.eluxemburgensia.lu/ark:70795/f869s2/pages/5/articles/DTL1060

[13] https://viewer.eluxemburgensia.lu/ark:70795/f869s2/pages/23/articles/DTL1114

[14] https://fr.wiktionary.org/wiki/grain

[15] https://en.wiktionary.org/wiki/grain

[16] https://viewer.eluxemburgensia.lu/ark:70795/q016qs/pages/3/articles/DTL609,

[17] https://viewer.eluxemburgensia.lu/ark:70795/q016qs/pages/17/articles/DTL631

[18] https://fr.wiktionary.org/wiki/grain

[19] https://en.wiktionary.org/wiki/grain

[20] Centre National de Ressources Textuelles et Lexicales. https://www.cnrtl.fr/portail.

[21] ChatGPT-4, via a subscription account, was prompted to select from the neighbours corresponding to a certain time slice the ones most likely to align with one of the senses of the term *révolution* listed by the lexical portal and attached to the prompt as a PDF file. A corpus citation was also added to the prompt to increase the sense alignment likelihood. For more details on the methodology and prompt examples, see Armaselu et al. (2024c).

Table 3: BnL French monographs: *révolution*, selected neighbours, corpus context and CNRTL-Ortolang entries

Term	**1690-1794**	**1831-1866**	**1867-1889**	**1890-1918**
révolution	‘moyene’, ‘ajouter’, ‘chant’, ‘envelopper’, ‘tige’, ‘engrennat’, ‘chaussée’, ...	‘paraboloïde’, ‘polaire’, ‘lemniscate’, ‘ellipsoïde’, ‘cubature’, ...	‘écroulement’, ‘plutonien’, ‘explosion’, ‘nationalité’, ‘avènement’, ...	‘vandalisme’, ‘insurrection’, ‘insurgé’, ...
BnL monogr. 1690- 1918	« La roue de longue tige ou grand moyene fait une révolution par heure ... » F. Rosset, L’art de conduire et regler les pendules et les montres,[22] Veuve de J. B. Kleber, Luxembourg, 1789 (p. 13) [23]	« ... et alors on aura le paraboloïde de révolution.» J.-N. Noël, Mémoires de la Société royale des sciences de Liège,[24] H. Dessain, Liège, 1844 (p. 164) [25]	« ... grandes révolutions plutoniennes qui à l’époque de la formation, ont bouleversé ... » Albert Gras, Trois ans dans l’Amérique du Sud,[26] Jos. Beffort, Luxembourg, 1883 (p. 22) [27]	« ... échappés au vandalisme de la révolution française ... » Un membre de l’œuvre des missionnaires luxembourgeois, Histoire de Notre-Dame de Luxembourg, Librairie Erpelding, Luxembourg, 1904 (p. 10) [28]
CNRTL- Ortolang. révolution[30]	*Mec. Mouvement circulaire effectué par un corps autour de son axe ....* Att. 1727. (en. motion of a body around an axis)	*Geom. Mouvement effectué autour d’un axe de rotation immobile par une ligne, ... une figure mathématique.* Att. 1799. (en. motion of a figure around an axis)	*Géophys. ... phénomènes naturels ... qui ont marqué la surface de la terre.* Att. 1749. (en. natural phenomena)	*Hist. ... La Révolution française ...* Att. 1789. (en. the French Revolution)

One can observe that the different senses of *revolution*[31] (movement of a body or a mathematical figure, and natural or socio-political phenomena) appear as posterior to the CNRTL-Ortolang attestation dates. It should be noted that these dates refer to Belgian and Luxembourgish publishers. Further investigation may uncover possible interplays between the spatial and temporal dimensions, and

[22] https://viewer.eluxemburgensia.lu/ark:70795/dqgfr3/pages/5/articles/DTL592

[23] https://viewer.eluxemburgensia.lu/ark:70795/dqgfr3/pages/17/articles/DTL612

[24] https://viewer.eluxemburgensia.lu/ark:70795/0vh3mc/pages/5/articles/DTL1732

[25] https://viewer.eluxemburgensia.lu/ark:70795/0vh3mc/pages/234/articles/DTL1790

[26] https://viewer.eluxemburgensia.lu/ark:70795/n3234k/pages/5/articles/DTL1589

[27] https://viewer.eluxemburgensia.lu/ark:70795/n3234k/pages/34/articles/DTL1589.

[28] https://viewer.eluxemburgensia.lu/ark:70795/x8fmkm/pages/5/articles/DTL472

[29] https://viewer.eluxemburgensia.lu/ark:70795/x8fmkm/pages/16/articles/DTL472

[30] https://www.cnrtl.fr/definition/r%C3%A9volution

[31] *Révolution*: lemma frequency for the whole corpus: 472; time slice frequency: 1690-1794: 16; 1831-1866: 276; 1867-1889: 97; 1890-1918: 82.

the circulation of cultural artefacts and of knowledge between these countries at the time. For instance, words such as *maltine*, *blédoline* (table 2), appear in none of the two dictionaries. Are they indicating linguistic innovations or forms with limited circulation that were not recorded and attested in standard dictionaries?

While Ortolang provides attestation information in a more systematic way, Wiktionary provides cross-language links for some etymologies. *Grain* and *révolution* are therefore linked to the Latin *granum* and *revolutio*. Ortolang etymology lists, however, the development of their various forms and meanings in French together with attestation dates and citations. Wiktionary offers other types of cross-lingual relations, such as translations in various languages, while both dictionaries display relations of synonymy, antonymy (Ortolang), or hyponymy and derivation (Wiktionary).

Another lemma that we analysed was *ville*[32] originating from the Latin *villa* (en. city). The neighbours included lemmas in older forms such as *chastellenie* as in « la ville et chastellenie de Marville ».[33] The corpus also included the modern form *châtellenie* as in « la restitution de la châtellenie de Ligny » [34] (en. castellany, jurisdiction of a castellan). Ortolang attests the modern form in 1740. This illustrated another aspect of the evolution of a concept, the change of the form of a term without the change of meaning.

2.2. LatinISE

In LatinISE [35] (McGillivray and Kilgarriff 2013) (Figure 2), the majority of texts were extracted from the IntraText[36] digital library. The corpus was automatically lemmatised and PoS-tagged. The annotation was partially corrected by

[32] *Ville*: lemma frequency for the whole corpus: 5713; time slice frequency: 1690-1794: 9; 1867-1889: 2113.

[33] J. Schœtter, Luxembourg et le comté de Chiny depuis le traité de paix de Nimègue jusqu'à la prise de la ville de Luxembourg par Louis XIV (1678 - 1684), https://viewer.eluxemburgensia.lu/ark:70795/mxtc21/pages/3/articles/DTL342 V. Bück, Luxembourg, 1880 (p. 19). https://viewer.eluxemburgensia.lu/ark:70795/gwd28x/pages/7/articles/DTL3346

[34] Le Chevalier L'évêque de la Basse Moûturie, Itinéraire du Luxembourg germanique, ou voyage historique et pitoresque dans le Grand-Duché https://viewer.eluxemburgensia.lu/ark:70795/gwd28x/pages/7/articles/DTL3346, V. Hoffmann, Luxembourg, 1844 (p. XII https://viewer.eluxemburgensia.lu/ark:70795/gwd28x/pages/28/articles/DTL3346).

[35] https://www.sketchengine.eu/latinise-corpus/

[36] www.intratext.com.

hand. LatinISE can be queried in Sketch Engine[37] and is available as one single text file in the vertical format at the LINDAT/CLARIAH-CZ Repository.[38] We preprocessed LatinISE to exclude texts for which no date was available, and recorded the (approximate) date of creation for the remaining texts, with negative dates corresponding to BCE dates. We worked on the lemmatised version of the corpus, divided the texts into sentences (delimited by strong punctuation marks), and excluded all punctuation marks. A portion of the metadata is in Table 4.

To ensure sufficient data was available for each time slice, we divided the corpus into three intervals: from 450 BCE to 1 BCE, from 1 CE to 450 CE, and from 451 CE to 900 CE. The number of tokens, types and unique lemmas in each time slice is shown in Table 5. Following Sprugnoli et al. (2019) and Ribary et al. (2020), for each time interval we trained a fastText model (Bojanowski et al. 2017) with 100 dimensions, context window of 10, and minimum frequency count of 50.

Table 4: LatinISE. Structure of the metadata of the texts with some examples.

ID	Title	Creator	Date	Type
IT-LAT0001	Vulgata	Hieronymus	382	poetry
IT-LAT0537	Ars Amatoria	Ovidius Naso, Publius	-9	poetry
IT-LAT0011	S. Benedicti Regula	Benedictus Nursianus	524	prose

After training the embeddings on each time slice of the corpus, we aligned the embedding spaces using Orthogonal Procrustes (Schönemann 1966) and measured the cosine similarity between an embedding of a word in one time interval and the embedding of the same word in the last time interval.

Table 5: LatinISE (number of tokens, types, and unique lemmas for each time slice)

Time interval	Tokens	Types	Unique lemmas
450 BCE-1 BCE	1,395,858	103,432	44,861
1 CE-450 CE	2,799,762	195,764	97,396
451 CE-900 CE	1,105,116	97,905	50,265

[37] https://auth.sketchengine.eu/

[38] https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-3170

Building on (McGillivray and Nowak 2022), we focused on the following list of polysemous socio-political terms referring to changing institutions of ancient and early medieval society (Thomas 2012, Lyasse 2007, Lorenzo 1976): *civitas* 'citizenship; city', *consilium* 'resolution; judgement; council', *senatus* 'Senate; council', *hostis* 'foreigner; enemy'; *imperator* 'general; emperor', *natio* 'birth; nation', *pontifex* 'pontifex; bishop', *potestas* 'power; magistracy'.

Table 6: LatinISE. *Civitas, pontifex*, and *potestas*: top neighbours for each time slice.

Term	**450BCE-1BCE**	**1CE-450CE**	**451CE-900CE**
civitas	gens, libertas, legatio, servitus, societas, ditio, potens, potentia, princeps, status	urbs, oppidum, murus, porta, provincia, regio, domus, castellum, vasto, Italia	urbs, insula, villa, oppidum, castrum, regio, castellum, vicus, Gallia, sedes
pontifex	dedico, Iulius, A., Paulus, Cornelius, C., aedes, sacerdos, nepos, annalis	sacerdos, Paulus, praetor, designo, consul, Cornelius, praepositus, magistratus, Marcus, A.	beo, sacerdos, sedes, nuper, missa, beatus, Paulus, propero, princeps, civis
potestas	arbitrium, ditio, libertas, ius, maiestas, restituo, beneficium, jussus, iudicium, lex	arbitrium, ditio, maiestas, ius, petitio, societas, census, auctoritas, condicio, liberalitas	parens, liberi, ius, liber, facultas, servitus, filius, mater, naturalis, serva

Table 6 shows the top neighbours for the words *civitas, pontifex*, and *potestas* in the three time slices. The neighbours of *civitas* seem to shift from words indicating concepts around citizenship such as *gens* 'people' and *societas* 'society' to geographical terms such as *urbs* 'city', *oppidum* 'city', and *regio* 'region' in the second time interval. The neighbours of *pontifex* seem to have shifted from words related to the Pagan religion (as indicated by neighbours such as *sacerdos* 'priest' and *aedes* 'sanctuary') towards words related to Christianity, such as *missa* 'mass' and *beatus* 'blessed'. In the case of *potestas*, words related to the law domain (e.g., *arbitrium* 'judgement' and *iussus* 'order') are replaced by names of family members (e.g., *parens* 'parent', *liberi* 'children') in the last time interval, indicating potentially a shift from public to more private legal relationships.

2.3. Responsa

Responsa[39] (Figure 2) spans generations and has specific characteristics, as the following. (1) Since Jewish people were dispersed for two millennia, local languages affected spoken Hebrew, resulting in textual diversity. (2) The Hebrew language evolved over the course of a millennium, and the vocabulary and style of contemporary and ancient writings varied considerably. (3) A response document contains all the Jewish legal arguments that led to a decision. It references a variety of preceding sources, including the Talmud and its commentators, legal laws, and earlier responses. (4) The corpus comprises texts written in Hebrew, Yiddish, and Aramaic, as per (1) and (3). The corpus is a commercial product that contains edited texts without OCR errors. It has been utilised in diachronic analysis using word embeddings (Zohar et al. 2013, Liebeskind et al. 2016, Liebeskind and Liebeskind 2020).

The Responsa is organised into four periods that show the development of halakhic (religious law) cf. Figure 3. For each era, we built a word embedding model using word2vec (Mikolov et al. 2013, Řehůřek and Sojka 2010) (5 word window, 100 dimension vectors). The vocabulary size for the developed word embedding models for each era is 6,848,442, 13,332,064, 32,756,802, and 37,505,021 respectively.

Experiments conducted on the Responsa focused on a subset of words also discussed in other diachronic corpora covered in this section (see subsections: 2.1, 2.2 and 2.4). In the 11th - 16th century, the word אזרח (*citizen, civilian*) was used in its biblical sense, referring to a significant inhabitant with status. Words such as בחיר (*elite*) מלך (*king*), and תושב (*residence*) can be found among its top neighbours. However, during the second and third periods (16th - 19th century), it was less prevalent. In the 19th century, the term “citizen” began to be used in its contemporary sense, referring to a person who is born or continuously resides in a country and, as a result, has full legal rights and political obligations. In its top neighbours, you can find the words מדינית (*political*), להבחר (*to be elected*) and various nationalities, including קנדי (*Canadian*), בריטי (*British*), etc.

[39] https://www.responsa.co.il/default.aspx

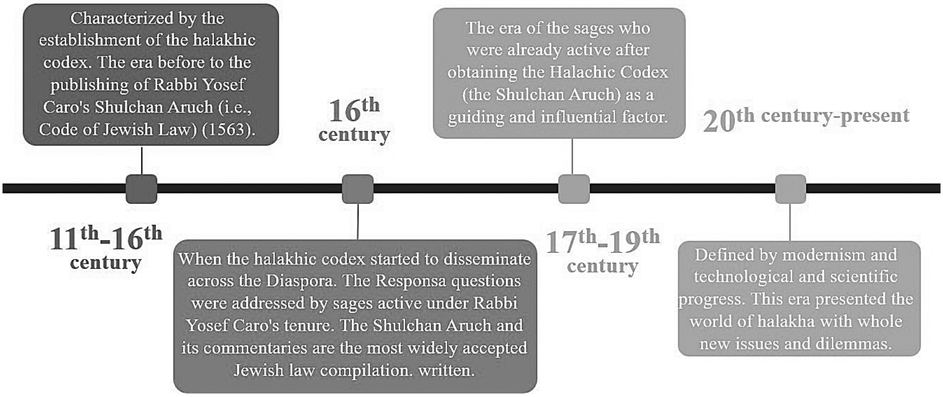

Characterized by the establishment of the halakhic codex. The era before to the publishing of Rabbi Yosef Caro's Shulchan Aruch (i.e., Code of Jewish Law) (1563).
16th century
The era of the sages who were already active after obtaining the Halachic Codex (the Shulchan Aruch) as a guiding and influential factor.
20th century-present
11th-16th century
When the halakhic codex started to disseminate across the Diaspora. The Responsa questions were addressed by sages active under Rabbi Yosef Caro's tenure. The Shulchan Aruch and its commentaries are the most widely accepted Jewish law compilation. written.
17th-19th century
Defined by modernism and technological and scientific progress. This era presented the world of halakha with whole new issues and dilemmas.

Figure 3: The four eras of the Responsa illustrate the historical evolution of halakhic ruling.

The Hebrew word בשורה (*gospel*) can alternatively be morphologically analyzed as ב+שורה (*in line*)[40], which appears frequently in the first era with terms such as מדרגה (*step*) and עמדה (*position*). Later, in the second and third eras, the frequent sense became *gospel* with top neighbours of (אבשר (*to apprise*), שדר (*to transmit*), אגרת (*letter*)) and (ברכות (*greeting*), נתבשר (*to be notified*), אבל (*mourner*)) respectively. Interestingly, during the fourth period, the sense of gospel predominates, but it is clear that the reference in the text to the Jewish custom of comforting mourners after the burial by making a line of comforters generates confusion between the two meanings.

Throughout the Responsa, the word מהפכה (*revolution*) has appeared in various contexts (as seen by its top neighbouring words). In the first era, references to revolution are mostly made in a religious context (כפירה (*atheism*), תשובה (*repentance*)). In the second era, the frequency of the word declines. However, in the third era, which corresponds to the periods in the French corpus, it appears in the context of war and tragedy (אונס (*rape*), הרג (*killing*), מיתה (*death*)) as a result of the pogroms and persecutions the Jews endured during this time. In the fourth era, industrial (מכונות (*machines*), אנרגיה (*energy*)) and medical (החיאה (*resuscitation*), אנאטומיה (*anatomy*)) revolutions as well as ideological revolution רפורמים (*(Judaism) Reform)*), חילוניות (*secularism*)) are addressed.

[40] Due to the variety of genres in the Responsa, existing Hebrew processing tools perform poorly on this corpus (Liebeskind et al. 2012). As a result, we did not do any morphological analysis.

2.4. Sliekkas

Sliekkas[41] (Gelumbeckaitė et al. 2012) (Figure 2) includes a representation layer which uses the original spelling, transliterated into modern Lithuanian on the next layer, followed by linguistic and morphological annotations. The text is lemmatised and English translations are provided. Additional Old Lithuanian corpora have been built in the Sliekkas tradition: CorDon (Drach 2021), a freely accessible, annotated corpus (ca. 24,000 words) of works by the Lithuanian national poet Kristijonas Donelaitis (1714–1780), and PosTime,[42] a 16th c. religious literature corpus developed since 2021.

These corpora were semi-automatically annotated using the Linguist's Toolbox (Buseman and Buseman 2013) and distributed in a tool-specific text format.

FastText is acknowledged as working better for word embeddings in morphologically rich languages with experimentally proven results in the Lithuanian language (Petkevicius and VitkuteAdzgauskiene 2021). For the research experiments, a first preliminary querying of Sliekkas for Lithuanian translation equivalents for the proposed words: *revolution / obole / civitas / potestas* was conducted.[43] The most frequent occurrences were also checked: *dievas (God)* - 573, *ponas (mister, lord)* - 251, *griekas (sin)* - 123, *evangelija (Gospel)* - 101. In further experiments *ponas (mister, lord)* and *griekas (sin)* were chosen as the semantic change was not likely to be expected in *dievas (God)* or *evangelija (Gospel)*. Concerning the semantics of *ponas (mister, lord)*, the Lithuanian language dictionary[44] provides the possible meanings which embrace (1) the privileged class, a rich person; (2) an independent person; (3) a title to address a man, and (4) an idle person. The word embeddings reveal that during the span of the 16th century, the word *ponas (mister, lord)* acquires the two meanings – a rich person and a title to address a man or possibly to stress the spiritual richness of Jesus as the most usual context observed is *ponas Jezus (Lord Jesus)* « bylojęs ponas Jėzus nė vienas negali dviem ponam slūžyti, tiem dviem, Dievui ir mamonui

[41] https://titus.fkidg1.uni-frankfurt.de/sliekkas/index.html

[42] https://gepris.dfg.de/gepris/projekt/443985248.

[43] Word occurrences: revolution/revolt (revoliucija/maištas) - 0, obole (grudai) - 15, civitas (miestas) - 23, potestas (valdžia) – 3.

[44] Lietuvių kalbos žodynas (Lithuanian language dictionary, electronic edition). 2017. Vilnius: Lietuvių kalbos insitutas. http://www.lkz.lt.

negalime slūžyti (said Lord Jesus, no one can break between two masters, we cannot break between those two, God and mammon) ». In the 18th-century span, the meaning of a rich person remains and the meaning of an independent person appears. The poetic text gives an example of a stork being the master of its own nest. « gandras ant savo lizdo - nei koks ponas išsisplėtęs (stork on his nest - like any gentleman spread) ». Concerning *griekas (sin)*, the Lithuanian Term dictionary provides only one meaning, that of a *sin*, and it is also noted that the word is an old borrowing which is currently not recommended to be used. In both time spans, it keeps the meaning of a *sin*.

2.5. RoDICA

For Romanian (Figure 2), static word embedding (word2vec) and contextual word embedding (ELMo) techniques (Truica et al. 2023) were applied to the RoDICA[45] corpus.[46] This corpus consists of Romanian-language news articles sourced from historical regions of Romania, i.e., Wallachia, Transylvania, Moldovia, and Bessarabia, and it encompasses a timeframe extending from the mid 19th century to the early 21st century. The RoDICA corpus was partitioned into four distinct subcorpora to facilitate regional analysis: RODICA-BS, encompassing texts originating from Bessarabia; RODICA-MD, comprising texts sourced from Moldavia; RODICA-TR, containing texts derived from Transylvania; RODICA-WL, consisting of texts collected from Wallachia. As a reference lexical resource, we used the online *Explanatory Dictionary of the Romanian Language – DEXonline*[47] that provided information regarding the etymology and the meaning of words.

In our experiments, we trained all the models on each of the 4 regional datasets. For the static embedding, we employed word2vec Skip-Gram with Negative Sampling (SGNS) with two training strategies. The first one, called SGNS-OP, trains the vectors in parallel on the time interval split corpora, while the second strategy, i.e., SGNS-WI, makes use of word injection. To capture the different meanings words can have depending on their context, we used ELMo (embed-

[45] https://relate.racai.ro/index.php?path=repository/resource&resource=rodica

[46] The code is available via https://github.com/DS4AI-UPB/SemanticChange-RO.

[47] https://dexonline.ro/

dings from language model) which is a Bi-LSTM-based language model. ELMo can create a more precise embedding that reflects the specific meaning used in each case, and this helps us better detect when the meaning of a word changes over time.

A diverse set of metrics were employed to detect semantic change in word meaning over time for this low resource language (Rosner et al. 2022). For instance, analyses focusing on local semantic neighborhoods are particularly effective at identifying culturally-driven semantic changes, while metrics assessing global displacement exhibit a greater sensitivity to more systematic and gradual shifts in word meaning. For the static word embedding models, we used the following distances: Euclidean distance, Manhattan distance, Canberra distance, Cosine distance, Bray-Curtis distance, and Correlation distance. To measure the performance of the contextual word embedding models, we used Average Pairwise distances, Jensen-Shannon divergence, and Cluster Count based on Affinity Propagation.

3. LLODIA model

Our tests with the multilingual datasets described in the previous section showed that after the detection of the neighbours, the interpretation process may involve the consultation of additional materials (dictionaries, historical sources) to understand the type of change observed in the embeddings. Semasiological and onomasiological perspectives on semantic change are already part of a major tradition in lexical semantics. Geeraerts (2010) considers that semasiological innovations (meaning-related) endow existing words with new meanings, while onomasiological innovations (naming-related) connect concepts to words in a way that is not yet part of a lexical inventory. The question that we will try to address is, therefore: can the LLOD formalisms be used to model such phenomena, and what type of reasoning may be envisaged to emulate the combination of corpus-based evidence and lexicographic information, and capture the complexity of the task in a multilingual diachronic context? The LLODIA model that we propose (Armaselu et al. 2024a, Armaselu et al. 2024b)[48] combines corpus and

[48] Model available via https://github.com/nexuslinguarum/LLODIA.

dictionary evidence on diachronic analysis and elaborates on LLOD formalisms such as OntoLex and OntoLex-FrAC (McCrae et al. 2017, Chiarcos et al. 2022a, Chiarcos et al. 2022b). For the modelling phase, we used the RDF-XML format and tools such as Oxygen XML Editor, Protégé and Vochbench for encoding, validation and querying the model.[49]

3.1. Classes and properties

Figure 4 shows a simplified LLODIA representation of the main classes and properties, as well as instances referring to the term *revolution* used as a proof of concept for our ideas. The class llodia:LexicalRecord was conceived as a recipient for recording information about linguistic events related to an ontolex:Form and the time slice[50] when such events were observed in a corpus. The record provides evidence about the frequency of the form observed in that time slice in the corpus and the values of the vector resulting from applying word embedding to that corpus segment, using properties such as frac:frequency and frac:embedding. A collection of records is represented by the class llodia:LexicalChronicle. We defined the invertible properties llodia:form, llodia:timeSlice and llodia:isRecordOf to link a record to a form, the corresponding time slice when the form is observed in the corpus, and the chronicle. The ontolex:LexicalConcept, to which a record may be connected via the llodia:lexicalConcept property, is used to define the meaning of the term expressed by the form through a list of neighbours (most similar words detected by cosine similarity)[51] and a frac:attestation relation that encapsulates attestation dates and citations from the corpus. If dictionary attestations and citations are available for a sense that can be aligned with the corpus evidence defined through the lexical concept, a connection to an ontolex:LexicalSense is also built. As shown in the figure, lexical senses can

[49] https://www.oxygenxml.com/. https://protege.stanford.edu/. https://vocbench.uniroma2.it/.

[50] Defined using the class dct:Period, with dct:start and dct:end time points, according to the Dublin Core specifications: http://purl.org/dc/terms/.

[51] The neighbour selection procedure slightly varied depending on the language and quality of the computed lists, e.g., semi-automatic selection using ChatGPT-4 from the top 20 most similar words (French), manual selection from the top 25-30 words (Hebrew), top 10 words with a frequency threshold of 5 (Latin). See also the descriptions attached to the frac:Similarity entries in the LLODIA proof of concept (GitHub repository).

be further connected to the ontolex:LexicalEntry that refers to the form under observation and which represents the target of the lexical record.

For the specific application of OntoLex-FrAC to linking concepts with corpus evidence, the class frac:Attestation was used. We utilised this class for attestations from dictionaries as well. For instance, to provide information about the subject assigned to a sense, e.g., (*Mechanics*), its definition or an explanation defining the meaning and an attestation quotation, including the attestation year, as provided by the lexicographical source. We created two separate classes, llodia:Corpus and llodia:Dictionary to distinguish these two types of resources and thus keep trace of their combination in analysing change in meaning over time. The property ontolex:reference was used to point to the corpus and dictionary resources consulted for the construction of the multilingual proof of concept.

We assume that temporal properties for attestation, such as start time and end time, as also suggested by Khan (2020), can be supported by the model to record situations similar to the *chastellenie* vs. *châtellenie* case, when the latter form replaced the former starting from 1740, with no change in meaning. Cross-lingual relations such as translations or etymologies in other languages were also represented in our model using vocabularies such as *lemonEty* (Khan 2018) and *vartrans*.[52] When this type of information was not available in the reference monolingual dictionaries, we used the data provided by Wiktionary.

We also encoded information about the corpus (source reference, publisher, title, language, publication date, time span) and the embeddings derived from the corpus (description of the applied method, quotation illustrating the usage of the sense, reference to the cited document, including its date of publication, and publisher). We used the classes frac:Frequency, frac:FixedSizeVector and frac:Similarity to store frequency values, word embeddings and lists of semantic neighbours corresponding to the observations of the target forms in the analysed corpora, and link these elements to lexical records and concepts, as previously explained at the beginning of this section.

[52] http://www.w3.org/ns/lemon/vartrans#, http://lari-datasets.ilc.cnr.it/lemonEty#.

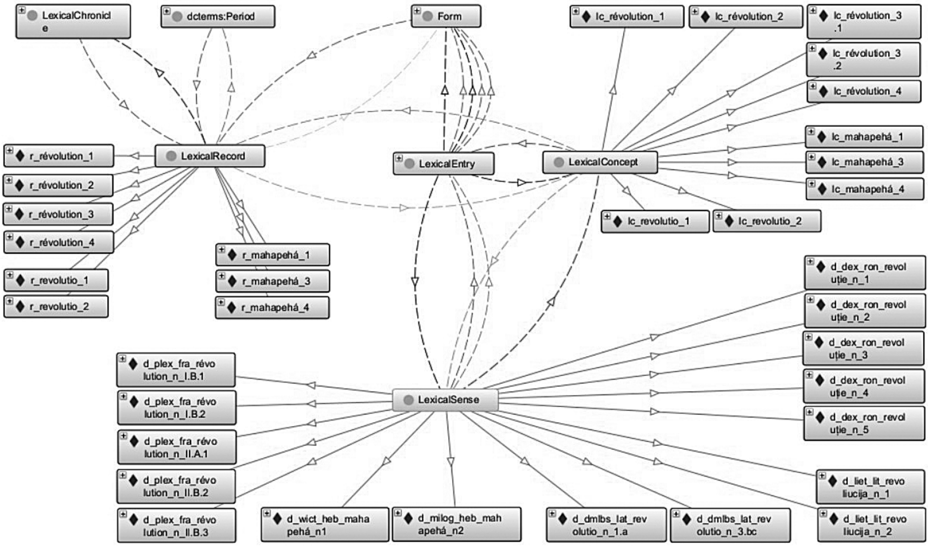


Figure 4: LLODIA classes and instances for the term *revolution* (Protégé)

We presume that this interconnected structure may allow for capturing both meaning-related and name-related changes in an observable entity. Innovation referring to new lexical forms with new meanings or unattested yet by the dictionaries should be possible to record when corpus evidence is identified (e.g., as in the case of *blédoline*, *maltine*, section 2.1).

It is important to note that the place of publication and of the provider or archiving source, if available, should also be included in the attestation data. Thus, the spatial and temporal dimensions may support discussions about potential cross-language and cross-cultural influences and knowledge circulation over time and space. The attestations can also include persistent links to digital facsimiles of the documents (when available), providing corpus evidence as proof of trusted sources, a requirement often evoked in historical research.

3.2. Instances

Figure 4 displays a set of instances for lexical records, concepts and senses corresponding to the term *revolution* in French, Hebrew, Lithuanian and Romanian, and the term *revolutio* in Latin, the etymon of all the others, except from the Hebrew *mahapehá* that has a different etymology. Since we lacked corpus evidence for Lithuanian and Romanian, lexical records and concepts were built only for the other three languages. However, with the information provided by the dictionaries, lexical senses were defined for all the five languages (with English used as a pivot for descriptions, encoding comments and metadata). This allowed us to build cross-language connections and query the model for aspects related to temporal and spatial dimensions, as showed in the following sections.

3.2.1. French

Table 3 displays the four time slices and selected neighbours for each of them in the BnL French monograph corpus corresponding to *révolution*. We modelled in LLODIA four lexical records corresponding to them, the one from the period 1867-1789 is discussed below. One can observe that the frequency of the form and the word embedding vector for the time slice and corpus segment are represented in the record:

```
<llodia:LexicalRecord rdf:about=”r_révolution_3”>
        <llodia:form rdf:resource=”f_révolution”/>
        <llodia:timeSlice rdf:resource=”ti_1867-1889”/>
        <frac:frequency rdf:resource=”freq_révolution_3”/>
        <frac:embedding rdf:resource=”fixedv_révolution_3”/>
        <llodia:lexicalConcept rdf:resource=”lc_révolution_3.1”/>
        <llodia:lexicalConcept rdf:resource=”lc_révolution_3.2”/>
</llodia:LexicalRecord>
```

While the other three lexical records of the term *révolution* were associated each to a single lexical concept, for this time slice two lexical concepts were identified. The two concepts were aligned to different dictionary senses from the CNRTL-Ortolang lexical portal, one related to the domain of “natural phenomena changing the physical characteristics of the Earth”, the other to the “sudden overthrow of the political regime of a nation.” The connection between the first pair lexical concept – lexical sense is illustrated by the following code:

```
<ontolex:LexicalConcept rdf:about=”lc_révolution_3.1”>
        <ontolex:reference rdf:resource=”c_bnlm_fra”/>
        <frac:embedding rdf:resource=”neighb_révolution_3.1”/>
        <frac:attestation rdf:resource=”ca_révolution_3.1”/>
        <ontolex:lexicalizedSense rdf:resource=”d_plex_fra_révolution_n_II.A.1”/>
</ontolex:LexicalConcept>
```

The selected neighbours (see also section 2.1) for the two concepts included terms such as *écroulement*, *plutonien*, *explosion* (collapse, Plutonian, explosion), for the first, and *nationalité*, *avènement*, *fédératif*, *insurgé* (nationality, advent, federative, insurgent), for the second. Examples 1 and 2 (Appendix, table 7) illustrate the corpus and dictionary attestations from 1883 and respectively 1749 that define the meaning of the target term as related to the domain of geology and natural phenomena. All the dictionary senses connected to a lexical concept were also related to a lexical entry, corresponding to the form under observation, as shown in figure 4 and example 3.

Similar modelling was devised for the other languages. The cross-language connections were built at the level of forms, both for translation and etymological relations. According to the reference dictionaries, the etymon of the term *revolution* in French, Lithuanian and Romanian is the Latin *revolutio*. Modelling examples and a discussion of its various meanings and attestations, as derived from the analysis of the Latin dataset and dictionary, are provided in the next section.

3.2.2. Latin

With reference to the *Dictionary of Medieval Latin from British Sources* (abbreviated as DMLBS) (Ashdowne 2016), accessed through the Logeion platform,[53] we established a sense inventory for the term *revolutio*, while relying on the LatinISE corpus to gather sense attestations. According to the DMLBS (Ashdowne 2016), *revolutio* has multiple senses, including: 1. rolling back or aside, 2. unrolling or opening a book, 3. circular movement or revolution, particularly in celestial or cyclical time references.[54] In the LatinISE corpus (McGillivray and Kilgarriff 2013), *revolutio* appears 21 times, solely in Medieval and early

[53] https://logeion.uchicago.edu/.

[54] The other senses are: 4. regular succession in office or rotation, 5. circular shapes like coils or spirals, 6. turning over, 7. reflection on past events, 8. repetition, and 9. relapse. Etymologically, it derives from the verb *revolvo*, meaning ‘to roll back’, ‘to unroll’, ‘to unwind’, or ‘to revolve’, with origins dating back to the Classical era.

modern texts. Two occurrences relate to sense 1, indicating rolling back or aside, specifically in stone movement contexts, which we modelled as follows:

```
<ontolex:LexicalSense rdf:about=”d_dmlbs_lat_revolutio_n_1.a”>
        <ontolex:reference rdf:resource=”d_dmlbs_lat”/>
        <rdfs:comment xml:lang=”eng”>(Act of) rolling back or aside.</rdfs:comment>
        <dct:subject rdf:resource=
                “https://dbpedia.org/page/Category:Motion_(physics)”/>
        <dct:source rdf:resource=”https://logeion.uchicago.edu/revolutio”/>
        <rdfs:comment rdf:resource=”lat_revolutio_n_1.a_assoc_descr”/>
        <frac:attestation rdf:resource=”da_revolutio_1”/>
</ontolex:LexicalSense>
```

The sentence in example 4 and its representation in LLODIA (Appendix, table 8) refers to the description of Jesus’ resurrection as it was found out by Mary Magdalene, according to the Gospel of John. In this passage, the word *revolutio* is used twice to describe the physical movement of the stone of the sepulchre (translated as ‘rolling away’ in the example). This corresponds to sense 1 in DMLBS.

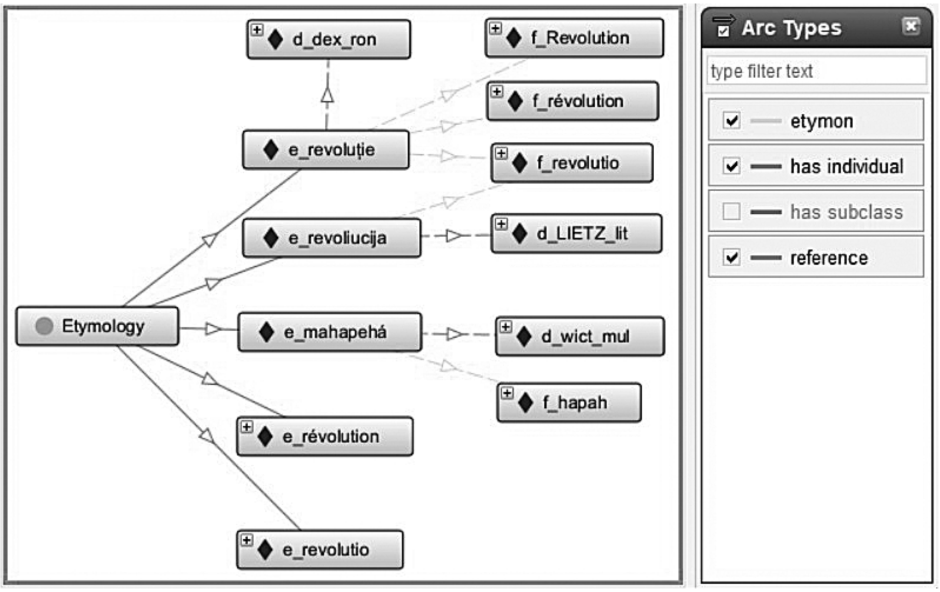
The remaining 19 instances of *revolutio* are instances of sense 3, referring to circular movement or revolution, particularly in celestial or cyclical time contexts. In example 5, Peter Abelard refers to the slaughter of newborns ordered by Herod as soon as he learned from the Wise Men that a new king was about to be born. In this passage, the word *revolutio* refers to the cycle of time corresponding to one year i.e., the time that Herod calculated for the baby to be born and order the mass slaughter (‘one year revolution’ in the example). This corresponds to sense 3 in DMLBS.

Given that the first attestations of the lemma with the two senses in the corpus are both from ca. 1000 CE, and that sense 1 only has two attestations (both in example 4), training the embeddings on different time spans does not lead to satisfactory results. Therefore, we trained fastText embeddings on the entire LatinISE corpus with a window size of 5 and a minimum frequency count of 5, deactivating the subwords option from the fastText model to exclude orthographically similar words from the closest neighbours. All of the closest neighbours (example 6) are related to the semantic field of astronomy, time calculation, or the rotational and revolutionary motion of the Earth around the sun.

None of them corresponds to the physical rolling motion illustrated in sense 1 of the DMLBS, which is understandable given the scarcity of occurrences of this sense within the corpus.

3.2.3. Hebrew, Lithuanian, Romanian

Similar encoding was produced for the three other languages, depending on the availability of data from the analysed corpora and reference dictionaries. Figure 4 shows three records for Hebrew, with three corresponding concepts derived from word embedding and two dictionary senses aligned with them. For Lithuanian and Romanian, no embedding results were available from corpora for the term *revolution*. Thus, we encoded the information provided by the dictionaries, for two and respectively five senses that were linked to lexical entries. Connections with the other languages included in the study were expressed through etymological and translation relations defined at the level of forms.



☑ etymon
☑ has individual
□ has subclass
☑ reference

Figure 5: LLODIA etymologies for *revolution* in Hebrew, Lithuanian and Romanian (Protégé)

Figure 5 displays the etymological relations between the Romanian form *revoluție* and its etymons in German, French and Latin, *Revolution*, *révolution*, *revolutio*. The Lithuanian form *revoliucija* is connected to the Latin *revolutio* while the Hebrew *mahapehá* shows a different etymon, *hapah*. All etymologies were encoded according to the information provided by the reference dictionaries for each language, also represented in the image. We used the class lemonEty:Etymology and defined the property llodia:etymon to specify the form that represents the etymon in the etymological description of another form.

3.3. Queries

In this section, we present a series of queries to demonstrate the functionality of our model, especially related to temporal, spatial and cross-language aspects.[55] The following SPARQL query run via Vocbench, illustrates how the model can be interrogated for attestation dates in certain time intervals.

```
SELECT DISTINCT ?attestation ?pub_place ?att_date ?subj ?expl WHERE {
        ?att rdf:type frac:Attestation .
        ...
        ?att dct:date ?att_date .
        ?cit rdf:type cito:Citation .
        ?att frac:citation ?cit .
        ?cit dbo:country ?pl .
        ?ls rdf:type ontolex:LexicalSense .
        ?lc rdf:type ontolex:LexicalConcept .
        {?ls frac:attestation ?att} UNION {?lc frac:attestation ?att . ?lc
        ontolex:lexicalizedSense ?ls .} ?ls dct:subject ?ls_subj .
        ?ls rdfs:comment ?expl.
        ...
        ?ls_att dct:date ?ls_att_date .
FILTER (((?att_date >= “1150” && ?att_date <= “1200”) || (?att_date >= “1790”
&& ?att_date <= “1830”) || (?att_date >= “1890” && ?att_date <= “1900”)) &&
LANG(?expl)=”eng”)}
```

[55] For additional examples of queries using LLODIA, see also (Armaselu et al. 2024b).

The five results provide information about the dictionary and corpus attestations (label starting with ‘da’ versus ‘ca’) for the five languages, the publication place, attestation year, the domain covered by the sense and a short explanation, in English, about the meaning.

```
Results count: 5
attestation pub_place att_date subj expl
“da_revolutio_2” “England” “1157” “Astronomy” “Act of revolving, circular movement,
        revolution (w. ref. to celestial motion and to cyclical passage of time).”@eng
“ca_mahapehá_1” “Israel” “1186” “Society” “Complete restitution, changing the existing
        order and habits.”@eng
“da_révolution_2” “France” “1799” “Geometry” “Motion of a geometric form around an
        axis.”@eng
“revoluţie_n_1” “Romania” “1821” “Politics” “Sudden, forcible overthrow of political power,
        causing fundamental changes in society.”@eng
“da_revoliucija_n_1” “Lithuania” “1894” “Politics” “Sudden, forcible overthrow of political
        power, causing fundamental changes in society.”@eng
```

The model can also be queried to provide the translation of the term *revolution* in English, used as a target language:

```
SELECT DISTINCT ?source ?target WHERE {
        ?trans_set rdf:type vartrans:TranslationSet .
        ?trans_set vartrans:source ?s_form.
        ?trans_set vartrans:target ?t_form.
        ?s_form rdf:value ?source .
        ?t_form rdf:value ?target .
FILTER (LANG(?target) = “eng”)}
```

The following results in the four other languages are obtained:

```
Results count: 4
source target
“révolution”@fra “revolution”@eng
“מהפכה mahapehá”@heb “revolution”@eng
“revoliucija”@lit “revolution”@eng
“revoluţie”@ron “revolution”@eng
```

4. Resource aggregator

The example modelled in the previous section should allow for further discussion about its potential applications in detecting and representing semantic change in a multilingual setting. Figure 6 illustrates such a scenario. As suggested by Hu et al. (2019), “change does not happen at a time point, but continuously throughout the process” (pp. 3899-3900). Our hypothesis is that we can mark specific milestones in the evolution of a concept on a timeline, by using attestation time, quotations, definitions, etc. provided by a lexicographical source such as a dictionary. For instance, if we consider a term and its equivalents (T1, T2, ..., Tn) in several languages (L1, L2, ..., Ln), it may be possible to identify the moments (t1, t2, ..., tp), when various senses (S11, S22, S1i, ...) of the term were attested by a dictionary or another type of reference document. With a resource as the one described in section 3, we can have access to complementary information from a corpus. The movement of the slider (red/darker vertical stick) on the timeline, can by analogy be associated to the consultation of available corpora to get a picture of the gradual changes that shaped the meaning of a term between the attestation of two different senses, e.g., S1i and S1m for language L1. An enquiry following the vertical red line may also inform us about this becoming through the lens of the different languages and cultures traversed by it in the study.

The LLOD modelling combining corpus- and dictionary-based resources therefore enables various forms of reasoning in the task of semantic change detection and representation within a multilingual context. We argue that a resource aggregator system making use of the flexibility and coverage of the Semantic Web paradigm can be helpful in capturing the interconnection between language changes and changes in the world (or extra-linguistic changes) (section 1). Figure 7 displays the basic modules of such a system.

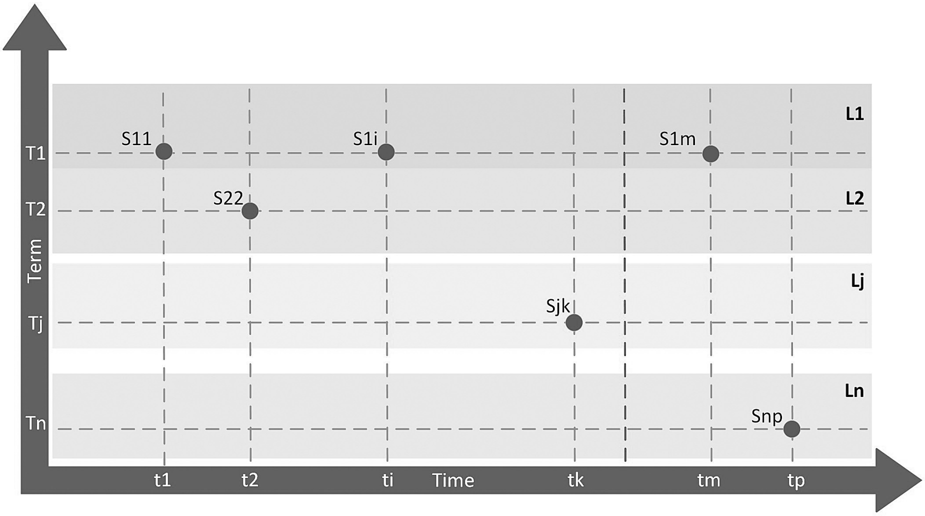


Figure 6: Multilingual semantic change timeline with attestation and query markers

Wiktionary and its OntoLex-based RDF edition DBnary (Sérasset 2015), can serve as an inspiration for creating such a resource. The LLODIA model (section 3) that joins curated and open collaborative resources and corpus evidence, can be used to query and compare new NLP results, such as word embeddings, with the already existing aggregated content.

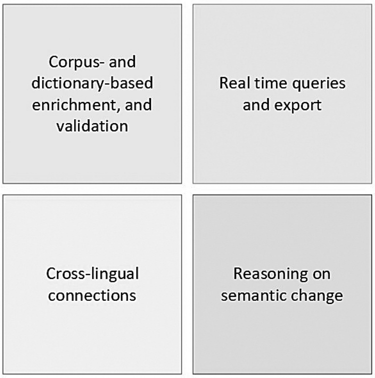

Corpus- and dictionary-based enrichment, and validation
Real time queries and export
Cross-lingual connections
Reasoning on semantic change

Figure 7: LLOD resource aggregator for tracing semantic change

This should enable reasoning about semantic change as illustrated through the experiments (section 2). For instance, cross-lingual inferences may be envisaged through the use of the etymology and translation relations, and queries that allow the researcher to trace parallel and sequential changes in meaning, polysemous similar or dissimilar behaviour, and possible connections between languages. See the examples of *revolution* in French, Latin, Hebrew, Lithuanian and Romanian, *city* in French and Latin, *citizen* and *mister, lord* in Latin, Hebrew, and Lithuanian. The aggregator should also allow for collaborative enrichment of the resource itself, validation, and export of attestations, citations, or timelines tracing a concept history in several languages.

We have published LLODIA, together with a small-scale multilingual diachronic sample presented above (Figure 4) via an open-source repository (GitHub). Such a resource can be further enriched by the community and queried by applications that combine, for instance, knowledge and corpus-based representations to derive sense embeddings (Hu et al. 2019, Scarlini et al. 2020), a trend that has showed promise as a more nuanced approach to tracing semantic change. The use of large language models (LLMs) and generative AI (GenAI) agents for semantic change detection and modelling represent another potentially promising avenue to be explored. However, for the implementation of a resource aggregator that can be queried and enriched in real-time by the community, a dedicated infrastructure would be needed. A topic that needs further investigation beyond the boundaries of this study.

5. Conclusion and future work

Starting from the hypothesis that a combination of theoretical perspectives from the history of concepts, diachronic word embedding, and LLOD has the potential to support semantic change analysis, we presented a set of experiments with multilingual corpora to illustrate the methodology. We proposed a model inspired by OntoLex-FrAC but that considers both corpus and dictionary evidence, and a conceptual framework for a resource aggregator that can be built to support such a task and features intended to query, reasoning and enrichment in real-time by the community. The main limitations of our approach consisted in the heterogeneity of the datasets and the sparsity of data for certain corpora used for analysis. These limitations were, to a certain degree, balanced by the aim of the project, more oriented towards the demonstrative approach rather than the performance assessment of fully automated processing and analysis of large data collections.

Further work may imply testing of more advanced methods, such as contextual word and sense embedding, use of LLMs and GenAI techniques, additional cross-language experiments and connections with lexico-semantic resources and knowledge databases such as WordNet and DBpedia. The digital infrastructure needed for the implementation of a resource aggregator as the one proposed above requires further study of the necessary resources and evaluation of the feasibility of the project.

Authors' contribution

F.A. wrote the initial manuscript, designed the general LLODIA model, and carried out the data processing, analysis and modelling, and section writing for French. B.M. and P.M. designed and carried out the data processing, analysis and modelling, and section writing for Latin. C.L. designed and carried out the data processing, analysis and modelling, and section writing for Hebrew. G.V.O. designed and carried out the data processing, analysis and modelling, and section writing for Lithuanian. E.S.A. and C.O.T. designed and carried out the data processing, analysis and modelling, and section writing for Romanian. All authors critically reviewed and approved the final version of the manuscript submitted to the Journal.

References

Abromeit, Frank; Chiarcos, Christian; Fäth, Christian; Ionov, Maxim. 2016. Linking the Tower of Babel: modelling a massive set of etymological dictionaries as RDF. *Proceedings of the 5th Workshop on Linked Data in Linguistics (LDL-2016): Managing, Building and Using Linked Language Resources*. Portoroz, Slovenia. 11–19.

Armaselu, Florentina; Apostol, Elena-Simona; Khan, Anas Fahad; Liebeskind, Chaya; McGillivray, Barbara; Truica, Ciprian-Octavian; Utka, Andrius; Valūnaitė Oleškevičienė, Giedrė; Van Erp,˙ Marieke. 2022. LL(O)D and NLP perspectives on semantic change for humanities research. *Semantic Web* 13.6. Eds. Cimiano, Philipp; Bosque-Gil, Julia; Cimiano, Philipp; Dojchinovski, Milan. 1051–1080. doi.org/10.3233/SW-222848.

Armaselu, Florentina; Liebeskind, Chaya; Marongiu, Paola; McGillivray, Barbara; Valunaite Oleskeviciene, Giedre; Apostol, Elena-Simona; Truica, Ciprian-Octavian; Gifu, Daniela. 2024b. LLODIA: A Linguistic Linked Open Data Model for Diachronic Analysis. *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*. Eds. Chiarcos, Christian; Gkirtzou, Katerina; Ionov, Maxim; Khan, Fahad; McCrae, John P.; Ponsoda, Elena Montiel; Chozas, Patricia Martín. Torino, Italia: ELRA and ICCL. 1–10.

Armaselu, Florentina; Liebeskind, Chaya; Marongiu, Paola; McGillivray, Barbara; Valūnaitė Oleškevičienė, Giedrė; Apostol, Elena Simona; Truică, Ciprian-Octavian. 2024a. Linguistic Linked Open Data for Diachronic Analysis (LLODIA). Dataset. doi.org/10.5281/zenodo.11065197.

Armaselu, Florentina; Liebeskind, Chaya; Valunaite Oleskeviciene, Giedre. 2024c. Self-Evaluation of Generative AI Prompts for Linguistic Linked Open Data Modelling in Diachronic Analysis. *Proceedings of the Workshop on Deep Learning and Linked Data (DLnLD) @ LREC-COLING 2024*. Eds. Sérasset, Gilles; Oliveira, Hugo Gonçalo; Oleskeviciene, Giedre Valunaite. Torino, Italia: ELRA and ICCL. 86–91.

Ashdowne, R. 2016. Data in online version of the 'Dictionary of Medieval Latin from British Sources' (DMLBS). Ed. Ashdowne, R.

Basile, Pierpaolo; Cassotti, Pierluigi; Ferilli, Stefano; McGillivray, Barbara. 2022. A New Timesensitive Model of Linguistic Knowledge for Graph Databases. *Proceedings of the 1st Workshop on Artificial Intelligence for Cultural Heritage co-located with the 21st International Conference of the Italian Association for Artificial Intelligence (AIxIA 2022).* CEUR Workshop Proceedings, 69.

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

Buseman, Karen; Buseman, Alan. 2013. Field Linguist's ToolBox (Version 1.6. 1). SIL International. Dallas, USA.

Camacho-Collados, Jose; Pilehvar, Mohammad Taher. 2018. From Word To Sense Embeddings: A Survey on Vector Representations of Meaning. *Journal of Artificial Intelligence Research* 63. 743–788. doi.org/10.1613/jair.1.11259.

Chiarcos, Christian; Apostol, Elena-Simona; Kabashi, Besim; Truica, Ciprian-Octavian. 2022a. Modelling Frequency, Attestation, and Corpus-Based Information with

OntoLex-FrAC. *Proceedings of the 29th International Conference on Computational Linguistics*. 4018–4027.

Chiarcos, Christian; Gkirtzou, Katerina; Ionov, Maxim; Kabashi, Besim; Khan, Fahad; Truică, Ciprian-Octavian. 2022b. Modelling Collocations in OntoLex-FrAC. *Proceedings of Globalex Workshop on Linked Lexicography within the 13th Language Resources and Evaluation Conference*. ERLA. 10–18.

Chiarcos, Christian; Silvano, Purificação; Damova, Mariana; Valunaite Oleškeviciene, Giedre; Liebeskind, Chaya; Trajanov, Dimitar; Truica, Ciprian-Octavian; Apostol, Elena-Simona; Baczkowska, Anna. 2023. Building an Owl-Ontology for Representing, Linking and Querying SemAF Discourse Annotations. *Rasprave Instituta za hrvatski jezik i jezikoslovlje* 49/1. 117–136. doi.org/10.31724/rihjj.49.1.6.

Chiru, Costin-Gabriel; Truica, Ciprian-Octavian; Apostol, Elena-Simona; Ionescu, Alexandru. 2021. Improving WordNet using Word Embeddings. *2021 23rd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*. IEEE. 121–128. doi.org/10.1109/ synasc54541.2021.00030.

Declerck, Thierry; Wandl-Vogt, Eveline; Mörth, Karlheinz. 2015. Towards a pan European lexicography by means of linked (open) data. *Proceedings of eLex*. 342–355.

Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Conference of the North American Chapter of the Association for Computational Linguistics*. 4171–4186. doi.org/10.18653/v1/N19-1423.

Drach, Mortimer. 2021. CorDon–A Deeply Annotated Digital Corpus of the Works of Kristijonas Donelaitis. *Archivum Lithuanicum* 23. 367–390.

Ehrmann, Maud; Romanello, Matteo; Clematide, Simon; Ströbel, Phillip Benjamin; Barman, Raphaël. 2020. Language Resources for Historical Newspapers: the Impresso Collection. *Proceedings of the 12th Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA). Marseille. 11.

Fokkens, Antske; Ter Braake, Serge; Maks, Isa; Ceolin, Davide. 2016. On the Semantics of Concept Drift: Towards Formal Definitions of Semantic Change. *Drift-a-LOD@ EKAW*. 10–17.

Gavin, Michael; Collin, Jennings; Kersey, Lauren; Pasanek, Brad. 2019. Spaces of Meaning: Conceptual History, Vector Semantics, and Close Reading. *Debates in the Digital Humanities 2019*. Eds. Gold, Matthew K.; Klein, Lauren F. University of Minnesota Press. 243–267. doi.org/10.5749/j.ctvg251hk.

Geeraerts, Dirk. 2010. *Theories of lexical semantics*. Oxford University Press. Oxford – New York.

Gelumbeckaitė, Jolanta; Šinkūnas, Mindaugas; Zinkevičius, Vytautas. 2012. “Senosios lietuvių kalbos tekstynas” (SLIEKKAS) - nauja diachroninio tekstyno samprata.

Darbai ir dienos 58. 257– 278.

Gromann, Dagmar et al. 2024. Multilinguality and LLOD: A survey across linguistic description levels. *Semantic Web* 15/5. 1915–1958. doi.org/10.3233/sw-243591.

Hu, Renfen; Li, Shen; Liang, Shichen. 2019. Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 3899–3908. doi.org/10.18653/v1/p19-1379.

Khan, Anas. 2018. Towards the Representation of Etymological Data on the Semantic Web. *Information* 9/12. 304. doi.org/10.3390/info9120304.

Khan, Anas Fahad et al. 2022. When linguistics meets web technologies. Recent advances in modelling linguistic linked data. *Semantic Web* 13/6. 987–1050. doi.org/10.3233/sw-222859.

Khan, Fahad. 2020. Representing Temporal Information in Lexical Linked Data Resources. *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020).* European Language Resources Association (ELRA). Marseille, France. 15–22.

Koselleck, Reinhart. 1994. Some Reflections on the Temporal Structure of Conceptual Change. *Main Trends in Cultural History. Ten Essays*. Eds. Melching, Willem; Velema, Wyger. Amsterdam Atlanta, GA: Editions Rodopi. 7–16.

Kutuzov, Andrey; Øvrelid, Lilja; Szymanski, Terrence; Velldal, Erik. 2018. Diachronic word embeddings and semantic shifts: a survey. *Proceedings of the 27th International Conference on Computational Linguistics*. 1384–1397.

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

Liebeskind, Chaya; Dagan, Ido; Schler, Jonathan. 2012. Statistical thesaurus construction for a morphologically rich language. * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. 59– 64.

Liebeskind, Chaya; Dagan, Ido; Schler, Jonathan. 2016. Semiautomatic construction of cross-period thesaurus. *Journal on Computing and Cultural Heritage (JOCCH)* 9/4. 1–26.

Liebeskind, Shmuel; Liebeskind, Chaya. 2020. Deep learning for period classification of historical Hebrew texts. *Journal of Data Mining & Digital Humanities* 2020. doi.org/10.46298/jdmdh.5864.

Lorenzo, Juan. 1976. Aportaciones al estudio léxico del latín de los cristianos. *Emerita* 44/2. 357– 371.

Lyasse, Emmanuel. 2007. Les rapports entre les notions de «res publica» et «ciuitas» dans la conception romaine de la cité et de l'Empire. *Latomus* 66/3. 580–605.

McCrae, John P; Bosque-Gil, Julia; Gracia, Jorge; Buitelaar, Paul. 2017. The OntoLex-Lemon Model: development and applications. *Proceedings of eLex 2017 Conference*. 587–597.

McGillivray, Barbara; Cassotti, Pierluigi; Di Pierro, Davide; Marongiu, Paola; Khan, Fahad; Ferilli, Stefano; Basile, Pierpaolo. 2023. Graph Databases for Diachronic Language Data Modelling. *Proceedings of the 4th Conference on Language, Data and Knowledge*. 86–96.

McGillivray, Barbara; Kilgarriff, Adam. 2013. Tools for historical corpus research, and a corpus of Latin. *New Methods in Historical Corpus Linguistics*. Eds. Bennett, Paul; Durrell, Martin; Scheible, Silke; Whitt, Richard J. Narr. Tübingen.

McGillivray, Barbara; Nowak, Krzyszof. 2022. Tracing the semantic change of socio-political terms from Classical to early Medieval Latin with computational methods. *Latin vulgaire – latin tardif XIV. 14th International Colloquium on Late and Vulgar Latin, September 5-9, 2022, Ghent University. Book of Abstracts.* Ghent University.

Mikolov, Tomas; Chen, Kai; Corrado, Greg; Dean, Jeffrey. 2013. Efficient estimation of word representations in vector space. *International Conference on Learning Representations*. 1–12.

OpenAI. 2023. *GPT-4 Technical Report*.

Peters, Matthew E.; Neumann, Mark; Iyyer, Mohit; Gardner, Matt; Clark, Christopher; Lee, Kenton; Zettlemoyer, Luke. 2018. Deep Contextualized Word Representations. *Conference of the North American Chapter of the Association for Computational Linguistics*. 2227–2237. doi.org/10.18653/v1/N18-1202.

Petkevicius, Mindaugas; Vitkute-Adzgauskiene, Daiva. 2021. Intrinsic Word Embedding Model Evaluation for Lithuanian Language Using Adapted Similarity and Relatedness Benchmark Datasets. *IVUS*. 122–131.

Řehůřek, Radim; Sojka, Petr. 2010. Software Framework for Topic Modelling with Large Corpora. *Workshop on New Challenges for NLP Frameworks*. 45–50.

Ribary, Marton; McGillivray, Barbara. 2020. A Corpus Approach to Roman Law Based on Justinian's Digest. *Informatics* 7/4.

Richter, Melvin. 1994. Begriffsgeschichte in Theory and Practice: Reconstructing the History of Political Concepts and Languages. *Main Trends in Cultural History. Ten Essays*. Eds. Melching, Willem; Velema, Wyger. Editions Rodopi. Amsterdam – Atlanta, GA. 121–149.

Rosner, Michael et al. 2022. Cross-Lingual Link Discovery for Under-Resourced Languages. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. ERLA. 181–192.

Scarlini, Bianca; Pasini, Tommaso; Navigli, Roberto. 2020. SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05. 8758–8765. doi.org/10.1609/aaai.v34i05.6402.

Schlechtweg, Dominik; McGillivray, Barbara; Hengchen, Simon; Dubossarsky, Haim; Tahmasebi, Nina. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. *arXiv:2007.11464 [cs].*

Schönemann, Peter H. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika* 31/1. 1–10.

Sérasset, Gilles. 2015. DBnary: Wiktionary as a Lemon-based multilingual lexical resource in RDF. *Semantic Web* 6/4. Eds. Hellmann, Sebastian; Moran, Steven; Brümmer, Martin; McCrae, John, 355–361. doi.org/10.3233/SW-140147.

Sprugnoli, Rachele; Passarotti, Marco; Moretti, Giovanni. 2019. Vir is to Moderatus as Mulier is to Intemperans - Lemma Embeddings for Latin. *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it)*. Bari, Italy.

Tahmasebi, Nina; Borin, Lars; Jatowt, Adam; Xu, Yang; Hengchen, Simon. 2021. *Computational approaches to semantic change*. Language Science Press. Berlin, Germany. doi.org/10.5281/ZENODO.5040241.

Thomas, Jean-François. 2012. Sur le champ lexical du pouvoir en latin. *Vita Latina* 185–186. 237– 249.

Tittel, Sabine; Bermúdez-Sabel, Helena; Chiarcos, Christian. 2018. Using RDFa to link text and dictionary data for medieval French. *Proceedings of the 6th Workshop on Linked Data in Linguistics (LDL-2016): Towards Linguistic Data Science.* European Language Resources Association (ELRA). Paris, France – Miyazaki, Japan.

Trajanov, Dimitar et al. 2024. From Linguistic Linked Data to Big Data. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).* ELRA and ICCL. 7489–7502.

Truică, Ciprian-Octavian; Tudose, Victor; Apostol, Elena-Simona. 2023. Semantic Change Detection for the Romanian Language. *2023 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC).* IEEE. 146–153. doi.org/10.1109/synasc61333.2023.00027.

Van Strien, Daniel; Beelen, Kaspar; Ardanuy, Mariona; Hosseini, Kasra; McGillivray, Barbara; Colavizza, Giovanni. 2020. Assessing the Impact of OCR Quality on Downstream NLP Tasks. *Proceedings of the 12th International Conference on Agents and Artificial Intelligence.* Valletta, Malta: SCITEPRESS - Science and Technology Publications. 484–496. doi.org/10.5220/0009169004840496.

Wang, Shenghui; Schlobach, Stefan; Klein, Michel. 2011. Concept drift and how to identify it. *Journal of Web Semantics First Look* 21.

Zohar, Hadas; Liebeskind, Chaya; Schler, Jonathan; Dagan, Ido. 2013. Automatic thesaurus construction for cross generation corpus. *Journal on Computing and Cultural Heritage (JOCCH)*

6/1. 1–19. doi.org/10.1145/2442080.2442084.

Višejezično ugrađivanje riječi i lingvistički povezani otvoreni podatci za praćenje semantičkih promjena

Sažetak

U članku se predlaže kombiniranje metoda obrade prirodnog jezika za dijakronijsku analizu i jezično povezanih modela otvorenih podataka za otkrivanje i predstavljanje semantičke promjene. Promjena značenja tijekom vremena riječi, fraza ili pojmova obuhvaća složene pojave koje se ne mogu u potpunosti objasniti samo distribucijskim metodama. Tvrdimo da pridruživanjem korpusnih i leksikografskih dokaza i modeliranjem rezultata u interoperabilnom formatu može pružiti čvršći temelj za donošenje zaključaka i mogućnosti ponovne uporabe u drugim aplikacijama. Definiramo osnovnu shemu za agregator resursa i model pod nazivom LLODIA (Jezicno povezani otvoreni podaci za dijakronijsku analizu). Da bismo ilustrirali naš pristup, koristimo višejezični skup podataka, na francuskom, latinskom, hebrejskom, starom litvanskom i rumunjskom jeziku, i gradimo uzorak izveden iz ugrađivanja riječi i resursa rječnika, kodiran pomoću predloženog modela.

Keywords: diachronic word embedding, linguistic linked open data, semantic change, multilingual datasets
Ključne riječi: dijakronijsko ugrađivanje riječi, lingvistički povezani otvoreni podatci, semantička promjena, višejezični skupovi podataka

Appendix

Table 7: LLODIA corpus and dictionary attestation and lexical entry examples (French)

Example	RDF-XML encoding
(1) *L’aspect général des Canaries est abrupt, de hautes falaises, continuellement battues par les eaux, en forment la base, et font réfléchir aux grandes révolutions plutoniennes qui à l’époque de la formation, ont bouleversé l’univers.* [The general aspect of the Canaries is abrupt, high cliffs, continually beaten by the waters, form its base, and make one think of the great Plutonian revolutions which, at the time of formation, turned the universe upside down.] (Albert Gras, *Trois ans dans l’Amérique du Sud. République de l’Uruguay. Aventures, chasses & mœurs*, 1883, pp. 22-23)	<frac:Attestation rdf:about="ca_révolution_3.1"> <ontolex:reference rdf:resource="c_bnlm_fra"/> <dct:date>1883</dct:date> <frac:citation> <cito:Citation rdf:about="cc_révolution_3.1"> <dct:title>Trois ans dans l’Amérique du Sud. République de l’Uruguay. Aventures, chasses &amp; mœurs</dct:title> <dct:creator>Albert Gras</dct:creator> <dct:publisher>Jos. Beffort, Succ. de L’Impr. Joris, Éditeur</dct:publisher> <dbo:country rdf:resource= "http://dbpedia.org/resource/Luxembourg"/> <rdf:value rdf:datatype="xsd:string"> L’aspect général des Canaries ... </rdf:value> <rdfs:comment>pp. 22-23</rdfs:comment> <dct:source rdf:resource= "https://viewer.eluxemburgensia.lu/ark:70795/ n3234k/pages/34/articles/DTL1599"/> </cito:Citation> </frac:citation> </frac:Attestation>
(2) The CNRTL-Ortolang lexical portal provides the following type of information for the attestation of this sense: *1749 géol. « phénomènes naturels qui ont bouleversé la surface terrestre »* [natural phenomena that have disrupted the Earth’s surface] (Buffon, Hist. et théorie de la terre, p. 96: *de grandes révolutions sur la surface de la terre* [great revolutions on the surface of the Earth]).	<frac:Attestation rdf:about="da_révolution_3.1"> <ontolex:reference rdf:resource="d_plex_fra"/> <dct:date>1749</dct:date> <frac:citation> <cito:Citation rdf:about="dc_révolution_3.1"> <dct:title>Hist. et théorie de la terre</dct:title> <dct:creator>Buffon</dct:creator> <dbo:country rdf:resource= "http://dbpedia.org/resource/France"/> <rdfs:comment>p. 96</rdfs:comment> <dct:source rdf:resource= "https://www.cnrtl.fr/definition/ r%C3%A9volution"/> </cito:Citation> </frac:citation> </frac:Attestation>
(3) Lexical entry corresponding to the form *révolution* and the dictionary senses from the CNRTL-Ortolang lexical portal aligned with the concepts identified by word embedding in the BnL Open Data corpus cut into time slices for diachronic analysis.	<ontolex:LexicalEntry rdf:about="le_révolution"> <ontolex:canonicalForm> <ontolex:Form rdf:about="f_révolution"> <ontolex:writtenRep xml:lang="fra"> révolution</ontolex:writtenRep> </ontolex:Form> </ontolex:canonicalForm> <lexinfo:partOfSpeech rdf:resource= "http://www.lexinfo.net/ontology/2.0/lexinfo#noun"/> <ontolex:sense rdf:resource="d_plex_fra_révolution_n_I.B.2"/> <ontolex:sense rdf:resource="d_plex_fra_révolution_n_I.B.1"/> <ontolex:sense rdf:resource="d_plex_fra_révolution_n_II.A.1"/> <ontolex:sense rdf:resource="d_plex_fra_révolution_n_II.B.2"/> <ontolex:sense rdf:resource="d_plex_fra_révolution_n_II.B.3"/> </ontolex:LexicalEntry>

Table 8: LLODIA corpus attestation, lexical sense and similarity examples (Latin)

Example	**RDF-XML encoding**
(4) *Quod cum minime reperisset, iterum ad monumentum orto iam sole cum aliis venit, et tunc revolutio lapidis facta est, quamvis Ioannes hanc revolutionem quasi prius factam per anticipationem dicat a Maria visam fuisse.* [As she (i.e., Mary Magdalene) had found out very little about it (i.e., about Jesus Christ's resurrection), after the sunrise she went up again with others to the sepulchre, and then the rolling away (i.e., of the stone) happened, although John says by anticipation that this rolling away had been seen by Mary as if it had happened earlier.] (Peter Abelard, *Problemata Heloissae cum Petri Abaelardi solutionibus* 5)	<frac:Attestation rdf:about="ca_revolutio_1"> <ontolex:reference rdf:resource="c_latinise_lat"/> <dct:date>1110</dct:date> <frac:citation> <cito:Citation rdf:about="cc_revolutio_1"> <dct:title>Problemata Heloissae </dct:title> <dct:creator>Petrus Abaelardus </dct:creator> <dbo:country rdf:resource= "http://dbpedia.org/resource/France"/> <rdf:value rdf:datatype="xsd:string"> Quod cum minime reperisset ... </rdf:value> <dct:source rdf:resource= "http://hdl.handle.net/11372/LRT-3170"/> </cito:Citation> </frac:citation> </frac:Attestation>
(5) *Si ergo Herodes ab eo quod didicit a magis stellam apparuisse biennium computaverit, plus quam integrum annum depressit de tempore, cum videlicet constet hanc occasionem nequaquam peragi, nisi post anni revolutionem quam nos hodie huius interfectionis passionem recolimus.* [Therefore, if Herod calculated two years from the moment that he learned from the Wise Men that the star had appeared, he subtracted more than one year from time, as it is well established that by no means this opportunity would be accomplished until after an year's revolution, which today we recall as the suffering of his killing.] (Peter Abelard, *Sermones* 34)	<ontolex:LexicalSense rdf:about= "d_dmlbs_lat_revolutio_n_3.bc"> <ontolex:reference rdf:resource="d_dmlbs_lat"/> <rdfs:comment xml:lang="eng">Act of revolving, circular movement, revolution (w. ref. to celestial motion and to cyclical passage of time). </rdfs:comment> <dct:subject rdf:resource= "https://dbpedia.org/page/Category:Astronomy"/> <dct:source rdf:resource= "https://logeion.uchicago.edu/revolutio"/> <frac:attestation rdf:resource="da_revolutio_2"/> <rdfs:comment rdf:resource= "lat_revolutio_n_3_assoc_descr"/> </ontolex:LexicalSense>
(6) The ten nearest neighbors of *revolutio* in the model, along with their cosine similarity scores, were: *vergiliarum* 'Pleiades' (constellation) (0.80), *solstitialis* 'of the summer solstice or pertaining to solar revolution' (0.80), *autumnale* 'autumnal' (0.79), *solstitium* 'solstice' (0.78), *arcticum* 'northern, arctic' (0.77), *tricesima* 'the thirtieth' (0.77), *cente(n)simus* 'the hundredth' (0.77), *semicirculus* 'half-circle' (0.77), *sexdecim* 'sixteen' (0.76), and *octobri* 'of October' (0.76).	<frac:Similarity rdf:about="neighb_revolutio_2"> <dct:description rdf:resource="lat_neighb_descr"/> <rdf:value rdf:datatype="xsd:string"> 0.7996742725372314, ...</rdf:value> <llodia:similarity> <rdf:Seq> <rdf:li>vergiliarum</rdf:li> <rdf:li>solstitialis</rdf:li> <rdf:li>autumnale</rdf:li> <rdf:li>solstitium</rdf:li> <rdf:li>arcticum</rdf:li> <rdf:li>tricesima</rdf:li> <rdf:li>cente(n)simus</rdf:li> <rdf:li>semicirculus</rdf:li> <rdf:li>sexdecim</rdf:li> <rdf:li>octobri</rdf:li> </rdf:Seq> </llodia:similarity> <frac:observedIn rdf:resource="c_latinise_lat"/> <lexinfo:termType rdf:resource="ontolex:Form"/> </frac:Similarity>