LLODIA: A Linguistic Linked Open Data Model for Diachronic Analysis

Florentina Armaselu, Chaya Liebeskind, Paola Marongiu, Barbara McGillivray, Giedre Valunaite Oleskeviciene, Elena-Simona Apostol, Ciprian-Octavian Truică, Daniela Gifu

University of Luxembourg, Jerusalem College of Technology, Université de Neuchâtel, King's College London, Mykolas Romeris University, National University of Science and Technology Politehnica Bucharest, Romanian Academy - Iasi Branch

florentina.armaselu@uni.lu, liebchaya@gmail.com, paola.marongiu@unine.ch, barbara.mcgillivray@kcl.ac.uk, gvalunaite@mruni.eu, {elena.apostol, ciprian.truica}@upb.ro, daniela.gifu@uaic.ro

Abstract

This article proposes a linguistic linked open data model for diachronic analysis (LLODIA) that combines data derived from diachronic analysis of multilingual corpora with dictionary-based evidence. A humanities use case was devised as a proof of concept that includes examples in five languages (French, Hebrew, Latin, Lithuanian, and Romanian) related to various meanings of the term *revolution* considered at different time intervals. The examples were compiled through diachronic word embedding and dictionary alignment.

Keywords: linguistic linked open data, diachronic analysis, multilingual word embeddings

1. Introduction

In this article, we propose a model and dataset that bring together two areas of research often considered separately, linguistic linked open data (LLOD) and diachronic word embedding. The goal is to address the question of how to approach semantic change detection and modelling by combining algorithmic processing with the expressive power of the Semantic Web formalism (Khan et al., 2022). While our model and proof of concept were intended to represent the meaning of words based on corpus and dictionary evidence, they also served as a testbed for our ideas and a way of encoding through structured forms not only the analysis results but also our own understanding of how words and concepts evolve across language, time and space. The model called LLODIA (linguistic linked open data for diachronic analysis) elaborates on existing vocabularies and methods, such as OntoLex-Lemon, OntoLex-FrAC and the "perdurantist" approach (McCrae et al., 2017; Chiarcos et al., 2022a,b; Welty et al., 2006), and creates wrappers and bridges between concepts and resources previously not linked within the diachronic analysis context.

We started from the assumption that embedding results from semantic change analysis need to be assessed in a unified view against a reference background. For this purpose, we included in our modelling both information resulting from corpus processing and comparison with dictionary attestations. Our tests mainly consisted of static word embedding, gensim word2vec (Mikolov et al., 2013; Rehurek and Sojka, 2010) and fastText (Bojanowski et al., 2017), applied to our corpora in five languages (French, Hebrew, Latin, Lithuanian and Romanian). Experiments with contextual word embedding implementations such as AllenNLP (Gardner et al., 2018) and ELMo (Peters et al., 2018) have been applied so far to the Romanian corpus (Truică et al., 2023).

This paper focuses on the design of the LLODIA model and proof of concept. Section 2 presents the methodology devised for the different corpora in our dataset to build the model and the steps in the construction of the model itself. Section 3 explains in more detail the main LLODIA classes and properties and how the word embedding results have been modelled using them. In Sections 4 and 5, we discuss modelling examples and queries to illustrate the usage of the model. Section 6 synthesises our findings and presents some hypotheses for future work.

2. Methodology

Our method consisted of integrating diachronic word embedding results into LLOD modelling and including dictionary- and corpus-based evidence that referred to word meanings observed or attested at certain time points and intervals.

2.1. Diachronic Word Embedding

The French dataset contained a selection of about 6.4 million tokens from the National Library of Lux-

LDL: Linked Data in Linguistics @ LREC-COLING 2024, pages 1–10
20-25 May, 2024.

embourg Open Data monograph collection,[1] with a time span from 1690 to 1918. We cut the corpus into 6 time slices that were chosen based on events and periods related to the history of Luxembourg and the rules and policies regarding the use of the three languages (French, German, Luxembourgish) in the Grand Duchy.[2] These elements were considered to have an impact on the evolution of language and word meanings. The corpus was lemmatised and stopwords were removed. We applied gensim word2vec (100-dimension vectors, 5-word context window) to each time slice and cosine similarity measures to compute lists of neighbours for words belonging to topics such as socio-political, cultural, and historical. The word "révolution" was chosen for LLOD modelling since the different meanings detected and its potential for cross-language analysis were considered relevant to the study. The lexicographic resources used as references were the CNRTL's lexical portal[3] and Wiktionary.[4] The former offered rich attestation and etymological information about the analysed term. The latter provided multilingual information regarding etymology and translation in the five languages and English that we used as a pivot.

The Hebrew dataset comprised 76,710 articles, approximately 100 million word tokens sourced from the Responsa Project[5], spanning from the 11th century to the 21st century. The corpus was divided into four time periods, namely the 11th century until the end of the 15th century, the 16th century, the 17th through the 19th centuries, and the 20th century until the present day. These time periods were selected based on the historical development of halakhic (Jewish religious laws) rulings (Liebeskind and Liebeskind, 2020). These advancements were deemed to influence the evolution of language and the meanings of words. The Hebrew Responsa data set underwent minimal pre-processing before being used with gensim word2vec. The word2vec model used 100-dimensional vectors and a context window of 5 words. Due to the underwhelming performance of modern Hebrew POS taggers on the Responsa dataset (Liebeskind et al., 2012), the pre-processing step only involved tokenizing the text based on white spaces. The lexical resources utilized were Wiktionary and Milog[6]. The latter provided an additional meaning of the explored word that is present in the dataset but was not included in Wiktionary.

For the experiments on Latin we used LatinISE (McGillivray and Kilgarriff, 2013), a 13-million token corpus of Latin texts spanning from the 4th century BCE to the 21st century CE. We worked on the lemmatised version of the corpus. We trained a fastText model (Bojanowski et al., 2017) on LatinISE with 100 dimensions, a context window of 5, and a minimum frequency count of 5. We used the *Dictionary of Medieval Latin from British Sources* DMLBS (Ashdowne, 2016), accessed via the Logeion platform[7] to build a sense inventory for *revolutio*, and the LatinISE corpus to retrieve the sense attestations.

For the modelling experiments in Lithuanian, we used Sliekkas (Gelumbeckaitė et al., 2012) where the representation of the original spelling is transliterated into modern Lithuanian, followed by linguistic and morphological annotations. The lemmatised text was used for modelling from a freely accessible, annotated corpus (ca. 350,000 words) including 16th century religious literature and works by the Lithuanian national poet Kristijonas Donelaitis (1714–1780). Also for the sense attestations we used Lietuvių kalbos žodynas[8] and to identify the etymology, we referred to LIETUVIUZODYNAS.lt.[9]

To detect semantic change in Romanian, a low resource language, Truică et al. (2023) used two static word embedding techniques on the RoDICA corpus.[10] The experimental results showed that Word2Vec Skip-Gram with negative sampling and Orthogonal Procrustes (SGNS-OP) and Word2Vec Skip-Gram negative sampling and Word Injection (SGNS-WI) perform well in detecting semantic change on small datasets, while contextual word embeddings such as ELMo work better on larger datasets and are not suited for languages where collecting a large dataset can be a problem. Previously, Gifu (2016a,b) used RoDICA corpus to analyse topics over time and diachronic similarity between cognate languages by statistical analysis of word distribution over epochs. For Romanian, the RoDICA corpus did not contain any relevant occurrence of the showcase word "revoluție" (eng. revolution), thus the modelling using LLODIA only focuses on dictionary data from the online Explanatory Dictionary of the Romanian Language – DEXonline[11]. DEXonline acts as a lexical resource that offers information regarding the etymology and the different meanings of the target word.

[1] Bibliothèque nationale du Luxembourg (BnL) Open Data MONOGRAPH TEXT-PACK: https://data.bnl.lu/data/historical-newspapers/.

[2] For instance, the invasion of Napoleonic troops (1795), the Congress of Vienna (1815), the Royal Decree (1834) stating the official languages, etc.

[3] https://www.cnrtl.fr/portail/.

[4] https://www.wiktionary.org/.

[5] https://www.responsa.co.il/.

[6] https://milog.co.il/.

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

[8] http://www.lkz.lt/.

[9] https://www.lietuviuzodynas.lt/terminai.

[10] http://lsplr.iit.academiaromana-is.ro/resources/detail/7/

[11] https://dexonline.ro/

2.2. LLOD Modelling

The LLOD modelling included three main phases. Given the potential of generative AI (GenAI) and large language models (LLMs) to produce outputs in various tasks, such as math problem solving, coding and creative writing, based on step by step prompting (Wei et al., 2023; Chen et al., 2023), a series of prompts have been designed in the early stage to model in RDF-XML a set of examples based on the French word embedding results and dictionary consultation. The aim was to assist the team with RDF-XML modelling when expert assistance was not available. Tests with several GenAI agents were performed and after considering preliminary results, ChatGPT (OpenAI, 2023; Bubeck et al., 2023) and Microsoft Copilot (Ortiz, 2023) were selected for this task.

The prompts in the first phase included several categories. For instance, asking the agents general questions about RDF-XML syntax, class and property generation (Copilot), or to extract examples from an OntoLex-FrAC article (Chiarcos et al., 2022a) and express them into RDF-XML (ChatGPT-4). The RDF-XML format was chosen since XML was more familiar to the members of the team from the humanities area and having less experience with the Turtle language. Another category contained instructions for RDF-XML encoding of (1) resources (corpus, dictionaries), citations and related metadata (title, creator, publisher, publication date, time span), (2) embedding results (vectors, frequency counts, neighbour lists), and sense discrimination and dictionary alignments derived from the French use case on the term *révolution*. The goal was to create templates that could be used for the modelling examples in the other languages of the project.

In the second phase, the results of these conversations were analysed and compared with existing LLOD vocabularies, knowledge repositories and models, such as Dublin Core, DBPedia, `ontolex`, `frac`, `lexicog`, `lexinfo`, `vartrans`, `lemonEty`.[12] Then, the observations based on the French examples were generalised taking into account the broader LLOD context to define the classes and properties of the LLODIA model. Oxygen XML Editor[13] and Protégé[14] were used for creating, editing and validating the classes, properties and instances of the OWL-based implementation of the model.

Once the ontology and the first examples for French were created, validated and tested using the two editors, in the third phase, the model was enriched with examples in the other languages included in the study, and further refined based on observations and exchanges derived from the encoding of the various cases and their particularities. The following section describes in more detail the main characteristics of the proposed model.

3. LLODIA model

The main class of the LLODIA model is `LexicalRecord`, a wrapper around an `ontolex:Form`, which contains temporal information on when certain linguistic events about the form were observed. For this purpose a time interval was devised using the `dct:Period`[15] class was devised, including `dct:start` and `dct:end` properties, to be associated with the record. `LexicalRecord` was conceived as a subclass of `frac:Observable` referring to entities about which a series of corpus- and dictionary-based observations can be documented.

Figure 1 shows the connections of the class `LexicalRecord` to other classes. For instance, the invertible LLODIA properties `form`, `timeSlice`, `lexicalConcept`, and `isRecordOf` link a record with a form, the time interval in which a series of observations were performed, a lexical concept and a lexical chronicle (collection of lexical records). As shown in the figure, the chronicle contains 9 record instances, including "r_révolution_1" about the French form *révolution*, its frequency observed in the time slice 1690-1794, and an associated `frac:FixedSizeVector` resulting from applying static word embedding to that corpus segment.

Listing 1: Lexical concept related to a lexical record.

```
<ontolex:LexicalConcept rdf:about="
    lc_révolution_1">
    <ontolex:reference rdf:resource="
        c_bnlm_fra"/>
    <frac:embedding rdf:resource="
        neighb_révolution_1"/>
    <frac:attestation rdf:resource="
        ca_révolution_1"/>
    <ontolex:lexicalizedSense rdf:
        resource="
        d_plex_fra_révolution_n_I.B.2"/>
</ontolex:LexicalConcept>
```

Further information about the form was encoded by means of the class `ontolex:LexicalConcept` associated with the lexical record. We considered that lists of

[12] http://www.w3.org/ns/lemon/ontolex#, http://www.w3.org/ns/lemon/frac#, http://www.w3.org/ns/lemon/lexicog#, http://www.lexinfo.net/ontology/2.0/lexinfo#. http://www.w3.org/ns/lemon/vartrans#, http://lari-datasets.ilc.cnr.it/lemonEty#.

[13] https://www.oxygenxml.com/.

[14] https://protege.stanford.edu/

[15] http://purl.org/dc/terms/.

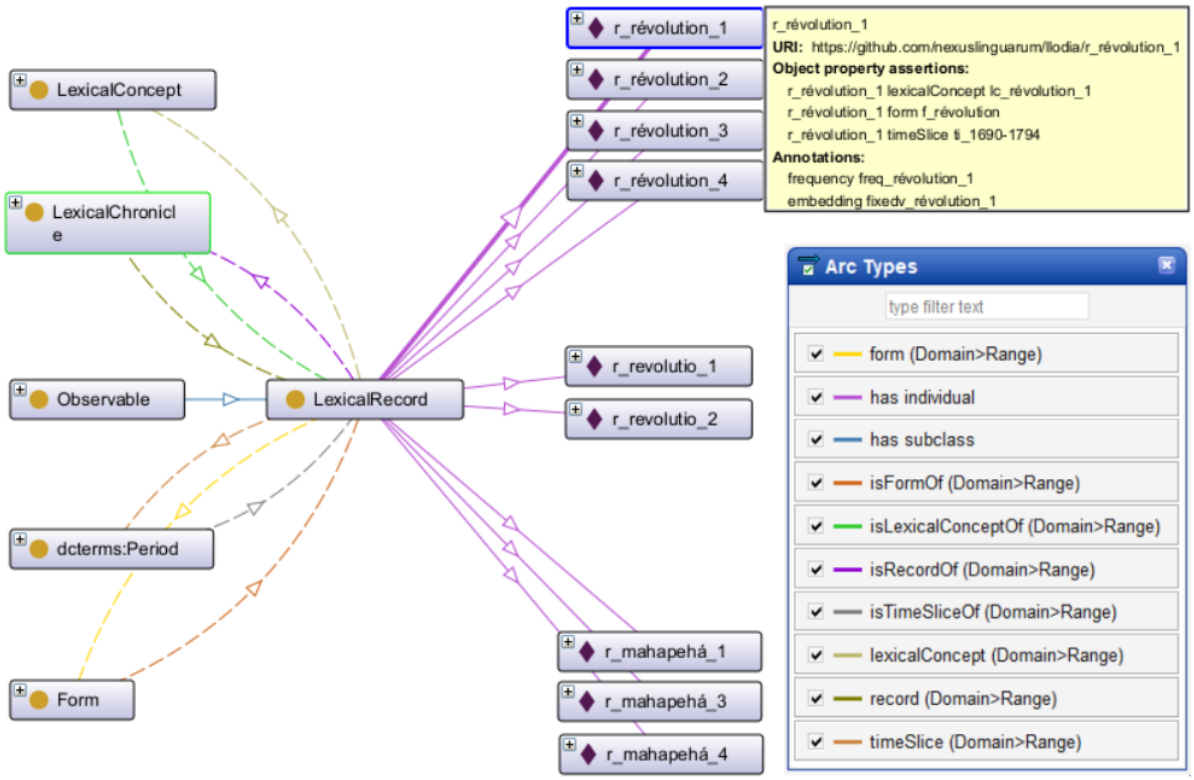

Figure 1: LexicalRecord and arc connections to classes (dashed) and individuals (solid) in Protégé.

neighbours and attestations from the corpus can capture certain aspects of the meaning of the word corresponding to the observed time period. Listings 1 and 2 describe the lexical concept associated with the "r_révolution_1" record, which refers to a list of neighbours computed through cosine similarity, a corpus attestation and a link to a lexical sense provided by a reference dictionary. Therefore, a `LexicalConcept` encompasses corpus-based evidence (neighbours, vector/embedding-related information, citation), while the related `LexicalSense` encapsulates dictionary-based evidence (sense, its domain and meaning explanation, attestation date). Thus, a record that documents the usage of a form observed in a certain time interval and corpus is indirectly connected through concepts to one or more senses in a reference lexicographical resource.

Listing 2: Corpus attestation of a lexical concept.

```
<frac:Attestation rdf:about="
    ca_révolution_1">
    <ontolex:reference rdf:resource="
        c_bnlm_fra"/>
    <dct:date>1789</dct:date>
    <frac:citation>
      <cito:Citation rdf:about="
          cc_révolution_1">
        <dct:title>L'art de conduire et
            regler les pendules ...
        </dct:title>
        <dct:creator>F. Rosset</dct:
            creator>
        <dct:publisher>Chez la Veuve de
            J. B. Kleber ...
        </dct:publisher>
        <dbo:country rdf:resource="http
            ://dbpedia.org/resource/
            Luxembourg"/> ...
        <rdf:value rdf:datatype="xsd:
            string">La roue de longue
            tige ou grande moyene fait
            une révolution par heure ...
        </rdf:value>
        <rdfs:comment>p. 13</rdfs:
            comment>
        <dct:source rdf:resource="https
            ://viewer.eluxemburgensia.lu
            /ark:70795/dqgfr3/pages/17/
            articles/DTL612"/>
      </cito:Citation>
    </frac:citation>
</frac:Attestation>
```

We defined two types of resources, `Corpus` and `Dictionary`, as LLODIA subclasses of `dcmitype:Collection` and `lexicog:LexicographicResource`. They were utilised to attest word forms and their meaning in a given time period and space, since information about the publishers and their location was also encoded, when available. Corpus

attestations were related to lexical concepts and associated neighbour list, while dictionary attestations were connected to lexical senses that were further linked to lexical entries corresponding to the observed forms (integrated as `ontolex:canonicalForm`). Translation and etymological relations across languages were encoded via `vartrans:TranslationSet` and `lemonEty:Etymology`, inspired by (Abromeit et al., 2016; Khan, 2018; Khan et al., 2020), and based on information extracted from multilingual resources such as Wiktionary or monolingual dictionaries. We considered that this type of corpus- and dictionary-based evidence allows the researcher to document and contextualise word meanings and their evolution and circulation over time and space.[16]

To test these assumptions, we created a set of interconnected examples for the term *revolution* in the six languages included in the study, with English as a pivot for general explanations of the sense meanings and descriptions of the process. When not enough evidence was available from the corpora, the information from the dictionaries was used instead. The following sections provide an overview of the observations encoded as a proof of concept and a series of queries on the model.

4. Multilingual Proof of Concept

The modelling task has drawn our attention to the dynamics of association between corpus and dictionary forms that express and record meaning characterisations and their usage over time and space. The following examples illustrate this aspect from the perspective of the datasets and languages considered for analysis.

4.1. French

The results of word embedding on the French corpus indicated that the term *révolution* occurred 16, 276, 97 and 82 times in four of the six time slices defined for analysis (1690-1794, 1831-1866, 1867-1889 and 1890-1918). For the neighbours intended to be included in the LLODIA encoding, we used the top 20 most similar words with *révolution* computed via cosine similarity. We devised a series of prompts for ChatGPT-4 to assist with the task of selection and alignment with dictionary senses. The agent was asked to separate the lists into sub-lists that could most likely be aligned with the senses of the word *révolution* according to the CNRTL's lexical portal. The process was iterative and in subsequent steps the citations extracted from the four corpus segments were also included in the prompts. Then, the output of the GenAI agent was manually checked and the terms from the sub-lists of neighbours considered most relevant to the chosen senses were selected.

The concept and associated dictionary sense for *révolution* assigned to the first time slice of the corpus corresponded to the domain of (1) mechanics as related to the circular motion of a body around its axis. The neighbours selected to model this concept included 10 terms, such as *moyene*, *ajouter*, *chant*, *envelopper*, *corde*, *tige*, with similarity measures between 0.89 and 0.79, and a citation from the field of clockwork mechanics describing the movement of wheels, minute and hour hands. The attestation date of this sense in the dictionary was 1727, with a citation from a French author, while the corpus citation was dated 1789 and indicated a Luxembourgish publisher. The list selected for the second corpus segment included 6 terms, e.g., *paraboloïde*, *polaire*, *lemniscate*, with similarity values between 0.65 and 0.58, and a citation pertaining to the domain of (2) geometry and the motion of a geometric form around an axis. The dictionary and corpus attestations pointed to the years 1799 and 1844, and to a French author and respectively Belgian publisher for the corpus citation.

A similar procedure was applied for the two other time intervals. The concepts and dictionary senses for *révolution* corresponding to them were related to the domains of (3) geophysics (natural phenomena changing the physical characteristics of the Earth) and (4) politics (sudden overthrow of the political regime of a nation) for the third segment, and (5) the French Revolution for the fourth one. The lists of neighbours selected for these concepts included terms such as *écroulement*, *plutonien*, *explosion* for concept (3), *nationalité*, *avènement*, *fédératif* for (4) and *vandalisme*, *insurrection*, *insurgé* for (5), with cosine similarity values in the range 0.70 - 0.61, 0.67 - 0.60 and respectively 0.64 - 0.57. The dictionary attestation years for the corresponding senses indicated 1749, 1636 and 1789, while the corpus attestation dates that we recorded for the related concepts were 1883 for (3) and (4), and 1904 for (5). GenAI prompts were tested for French, and then for Hebrew and Lithuanian and the outputs were manually checked and compared with the results of the evaluation method called LLM-Eval (Lin and Chen, 2023) applied for these languages.

4.2. Latin

According to the DMLBS (Ashdowne, 2016), the term *revolutio* has the following (main) senses: 1. (act of) rolling back or aside 2. (act of) unrolling or opening (book) 3. act of revolving, circular movement, revolution (referred to celestial motion or to cyclical

[16] The model and proof of concept has been published in the Nexus Linguarum GitHub repository (Armaselu et al., 2024): `https://github.com/nexuslinguarum/llodia/`.

passage of time); 4. regular and recurring succession of persons in office, rotation; 5. something that forms a circular shape, coil, spiral; 6. act of turning over 7. reflection on, consideration of, going back over a past event; 8. repetition 9. relapse. The term is etymologically derived from the verb *revolvo* which means 'to roll back; to unroll, unwind; to revolve, return' and is attested from the Classical era, e.g., in Cicero and Livius, although it becomes especially frequent in the Augustan period e.g., in Vergil.[17] In the LatinISE corpus (McGillivray and Kilgarriff, 2013), this lemma occurs 21 times, all in Medieval and early modern texts. It occurs twice, within the same sentence, in *Problemata Heloissae cum Petri Abaelardi solutionibus* by Peter Abelard (1110) with the sense 1 (act of rolling back or aside), referred to the movement of a stone.

The remaining 19 occurrences are found in the following texts, where *revolutio* expresses sense 3 (act of revolving, circular movement, revolution, referred to celestial motion or to cyclical passage of time): *Sermones* by Peter Abelard (1110); *De luce seu de inchoatione formarum* and *De impressionibus aeris seu de prognosticatione* by Robert Grosseteste (1200); *Missale Romanum* (1570). We trained fastText embeddings on LatinISE with window size 5 and minimum frequency count 5, turning subwords off.[18] The first ten closest neighbours of *revolutio* in the model (with their associated cosine similarity scores) are: *vergiliarum* 'Pleiades' (constellation)(0.80), *solstitialis* 'of the summer solstice or referred 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), *octobri* 'of october' (0.76). All the 10 closest neighbors refer to the semantic field of astronomy, time calculation, or the motion of rotation and revolution of the Earth around the sun. None of them pertains to the act of physical rolling motion i.e., the one illustrated in sense 1 in DMLBS. This is easily understandable given that this sense occurs only two times within the corpus, both in the same sentence, and therefore, the model training is affected by data sparsity.[19]

As it can be observed from the description of the occurrences of *revolutio* in the corpus, senses 1 and 3 are both attested for the first time in the corpus in 1110 (in the two texts by Peter Abelard). This, combined with the limited number of occurrences of *revolutio* with sense 1, has made it impossible to achieve satisfactory results when applying fastText on the corpus divided into smaller time spans.

4.3. Hebrew

Wiktionary defines the term in Hebrew מהפכה (revolution) as having the following meanings: 1. A historical event that significantly altered the trajectory of a specific nation or the course of human civilization as a whole. This could include revolutionary events like a technological revolution, such as the advent of the printing press, or a political upheaval like the French Revolution, which resulted in the overthrow of absolute monarchy. 2. Biblical terminology: destruction. 3. Derived from 2: chaos, commotion, a state of evident disarray. The Milog dictionary proposes an additional meaning for the word (4): Full restoration, altering the current arrangement and routines. The term has occurred in three distinct periods of the Responsa corpus (1st, 3rd, and 4th), each time in varying contexts. We obtained the 30 most closely related terms to the term מהפכה for each of the time periods. We manually chose 10 neighboring terms, excluding non-informative words that cannot be understood without context.

By examining the chosen terms, we assigned the most prevalent sense to each time period. The first period was assigned the fourth sense, as indicated by terms such as מהטעות (by the mistake)(0.72), החיסרון (the disadvantage)(0.71), and התועבה (the abomination)(0.698). These were primarily utilized in a religious context. The first sense has been assigned to the third and fourth periods. The third period is characterized by words such as ויהרגנו (and they killed us)(0.71), לאונסה (to rape her)(0.66) and שהזנות (that the prostitution)(0.65), which convey the themes of war and tragedy. This aligns with the historical periods of the French corpus, as it reflects the pogroms that Jews experienced during this period. The fourth period is characterized by neighboring words that are prominent in the context of medical and industrial revolutions, such as החייאה (resuscitation)(0.65), ממכונות (from machines)(0.646) and פאטולוגיה (pathology)(0.645).

It is important to observe that word2vec, as a non-contextualized approach, primarily provides terms that commonly occur in similar contexts as the given word. However, frequently, these contexts may not necessarily indicate the right sense of the word, even when used in the most prominent context. Moreover, on certain occasions, the word itself may be used in a manner that is outdated, conveying a meaning that is not explicitly defined in

[17] The entries for *revolvo* and *revolutio* are not yet available in the most comprehensive Latin lexicographic resource, the monolingual dictionary *Thesaurus Linguae Latinae* (Thesaurus-Kommission, 1900–), therefore we relied on the definitions and attestations provided in other Latin dictionaries.

[18] Tests with a higher minimum count and wider windows (10 to 50) led to unsatisfactory results. We turned subwords off in order to avoid getting ortographically similar words among the closest neighbours.

[19] Extending the number of closest neighbours to 20 did not improve the results.

the dictionary. For instance, a sentence extracted from the fourth period states: המכונה בעצמה כובסת הכביסה במה שהחשמל מהפכה וע``י זה נעשה הכביסה במכונה מעצמה ולא ע״י אדם (The machine itself washes the laundry as the electricity **turns** it and by this the washing is done in the machine by itself and not by a person). The context of this sentence is certainly related to an industrial revolution. However, the word מהפכה means turn which is not a direct sense of the word in the dictionary and is kind of archaic way to express the act of "turning" (הפיכה).

4.4. Lithuanian

For the modelling experiments related to the etymology of *revolution*, in Lithuanian we used the attestation of the dictionary LIETUVIUZODYNAS.lt which shows that *revoliucija* comes from Latin *revolutio*. Another dictionary, Lietuvių kalbos žodynas, identifies that the word was first mentioned in Lithuanian texts in the 19th century. Relying on the dictionary the word has two meanings: 1. *staigus prievartinis politinės valdžios nuvertimas, sukeliantis esminius visuomenės pakitimus (a sudden, forcible overthrow of political power, causing fundamental changes in society)*; 2. *kokybinis raidos pakeitimas (qualitative change of development)*.

4.5. Romanian

For the Romanian language, we used the DEXonline digital dictionary to determine the etymology and the different meanings of the word *revoluție* (en. revolution). According to this dictionary, the etymology of the word *revoluție* comes from three terms, i.e., the Latin term *revolutio*, the French term *révolution*, and the German term *Revolution*. The term *revoluție* has the following main senses: 1) a fundamental change in the values, political institutions, social structure, leaders, and ideologies of a society (in the philosophy field); 2) revolt, uprising; 3) a radical change or transformation in a certain field; 4) a continuous periodic motion of a body following a closed curve; 5) the rotational motion of a body around a fixed straight line (geometry); 6) the motion of a body that travels a fixed curve (physics); and 7) the geological change of the Earth's crust.

5. Queries

Once the conception of our model was stabilised and examples in all five languages were produced, we wanted to check the functionality of the LLODIA model through queries. For this purpose, we have chosen Vocbench[20] that included a SPARQL query editor.

Our intention was to test whether temporal aspects can be included in the queries to allow for time-based comparison across languages. Listing 3 illustrates how lexical records corresponding to a certain time interval can be retrieved from the model. In this case, four records, one from the Hebrew, and the other from the French dataset were retrieved.

Listing 3: Lexical record by time slice (Vocbench SPARQL).

```
PREFIX ...
SELECT DISTINCT ?lex_record ?t_start ?
    t_end WHERE {
    ?lrecord  rdf:type llodia:
        LexicalRecord .
    ...
    ?tslice rdf:type dct:Period .
    ?lrecord llodia:timeSlice ?tslice .
    ?tslice dct:start ?t_start .
    ?tslice dct:end ?t_end .
FILTER (?t_start >= "1600-01-01" && ?
    t_end <= "1900-12-31")}
Results count: 4
lex_record t_start t_end
"r_mahapehá_3" "1601-01-01" "1900-12-31"
"r_révolution_1" "1690-01-01"
    "1794-12-31"
"r_révolution_2" "1831-01-01"
    "1866-12-31"
"r_révolution_3" "1867-01-01"
    "1889-12-31"
```

Another element that seemed relevant to us in the context of diachronic analysis was the retrieval of attestation dates and places, to get an idea about when and where certain pieces of knowledge were produced. Listing 4 displays two dictionary and two corpus attestations, with their respective dates and place of publication for citations of the terms *revolutio* and *revolution* in Latin, Lithuanian, French and Hebrew and two time intervals.

Listing 4: Dictionary and corpus attestation by date and publisher place (Vocbench SPARQL).

```
PREFIX ...
SELECT DISTINCT ?attestation ?att_date ?
    pub_place WHERE {
    ?att rdf:type frac:Attestation .
    ...
    ?att dct:date ?att_date .
    ?cit rdf:type cito:Citation .
    ?att frac:citation ?cit .
    ?cit dbo:country ?pl .
    ...
FILTER ((?att_date >= "1150" && ?
    att_date <= "1180") || (?att_date >=
     "1890" && ?att_date <= "1920"))}
Results count: 4
attestation att_date pub_place
"da_revolutio_2" "1157" "England"
```

[20] https://vocbench.uniroma2.it/.

```
"da_revoliucija_n_1" "1894" "Lithuania"
"ca_révolution_4" "1904" "Luxembourg"
"ca_mahapehá_4_2" "1917" "Israel"
```

The query from listing 5 explores the possibility of finding similar domains across different languages, in which the various meanings of the retrieved terms were observed. The results display the domains of mechanics and astronomy and corresponding dictionary senses and their explanations in English for French, Latin and Romanian.

Listing 5: Sense by subject (Vocbench SPARQL).

```
PREFIX ...
SELECT DISTINCT ?lex_sense ?subj ?expl
    WHERE {
     ?ls rdf:type ontolex:LexicalSense .
     ?ls dct:subject ?ls_subj .
     ?ls rdfs:comment ?expl.
     ...
FILTER ((?subj = "Mechanics" || ?subj =
    "Astronomy") && LANG(?expl)="eng")}
Results count: 3
lex_sense subj expl
"d_plex_fra_révolution_n_I.B.2" "
    Mechanics" "Circular motion of a
    body around its axis."@eng
"d_dmlbs_lat_revolutio_n_3.bc" "
    Astronomy" "Act of revolving,
    circular movement, revolution (w.
    ref. to celestial motion and to
    cyclical passage of time)."@eng
"d_dex_ron_revoluție_n_3" "Mechanics" "
    Circular motion of a body around its
     axis."@eng
```

Translation relations can also be interrogated as illustrated in listing 6 that provides the translation of the French word *révolution* in English, Hebrew, Lithuanian and Romanian.

Listing 6: Translation (Vocbench SPARQL).

```
PREFIX ...
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(?source) = "fra")}
Results count: 4
source target
"révolution"@fra "revolution"@eng
"révolution"@fra "מהפכה mahapehá"@heb
"révolution"@fra "revoliucija"@lit
"révolution"@fra "revoluție"@ron
```

Finally, listing 7 presents a query about the etymons of the various forms stored in the model. The results show the common Latin root *revolutio* for *revolution* in French, Lithuanian and Romanian, the etymon of this root in Latin, and a different origin for Hebrew. Additional etymons are displayed for Romanian, the French form *révolution* and German *Revolution*. Etymological chains can be inferred, e.g., between the French, Lithuanian and Romanian forms, and the Latin *revolutio* and its etymon *revolvō*. It should be noted that both the translation and etymological relations were defined at the level of forms but other approaches, considering for instance connections at the sense level or complex etymological relations, can be imagined as well. These aspects are currently under study.

Listing 7: Etymology (Vocbench SPARQL).

```
PREFIX ...
SELECT DISTINCT ?form ?etymon
WHERE {
     ?frm rdf:type ontolex:Form .
     ?etm rdf:type lemonEty:Etymology .
     ?etym rdf:type ontolex:Form .
     ?frm lemonEty:etymology ?etm .
     ?etm llodia:etymon ?etym .
     ?frm rdf:value ?form .
     ?etym rdf:value ?etymon .
}
Results count: 7
form etymon
"révolution"@fra "revolutio"@lat
"מהפכה mahapehá"@heb "הפך hapah"@heb
"revoliucija"@lit "revolutio"@lat
"revolutio"@lat "revolvō"@lat
"revoluție"@ron "révolution"@fra
"revoluție"@ron "revolutio"@lat
"revoluție"@ron "Revolution"@deu
```

Our assumption was that this type of model can capture some of the complexities of the linguistic phenomenon of change in meaning over time and space, and across languages. Although the proof of concept contained a limited number of examples and was affected by data sparsity in some cases, it showed that interconnections can be built between time- and space-aware representations based on multilingual and varied types of resources. The sets of neighbours and corpus citations could provide insights into the contexts where a form occurred. The senses and attached domains could enable inferences about how the corresponding meanings, recorded by reference sources and reflecting the accepted usage by the community in a certain period of time, were possibly transmitted from one language to the other, evolved independently or influenced each other across linguistic and cultural borders, or disappeared.

6. Conclusion and future work

In this article, we proposed a LLOD model for diachronic analysis (LLODIA) and a proof of concept in five languages (French, Hebrew, Latin, Lithuanian, Romanian, with English as a pivot) for the term

revolution. We argue that a combination of corpus and dictionary evidence on the evolution of word meanings and its modelling in a structured format can provide a richer basis for analysing multilingual diachronic phenomena than each part alone. For this purpose, we used word embeddings computed on diachronic corpora, reference dictionaries and existing Semantic Web vocabularies, and created new classes and properties when the elements needed for our investigation were not available.

We used a set of queries to test the capabilities of the LLODIA model to express and support inferences based on time and space dimensions and interconnections across languages. While simple translation and etymological relations at the level of forms were considered at this stage, further enquiry is intended for more complex cases that require sense-level interrelations or etymological chains.

We designed LLODIA as a small-scale model and proof of concept that may serve as a starting point for other projects that combine NLP and LLOD methods to detect and represent change of meaning over time, space and across several languages. It can also be imagined as a larger lexicographic project based on interoperability with other vocabularies and expanded as an online resource aggregator that may be enriched, queried and reasoned upon by various contributors. However, for this type of interaction a dedicated infrastructure would be needed, which is a subject matter that needs additional study and examination.

7. Acknowledgments

This article is based upon work from COST Action *Nexus Linguarum, European network for Web-centred linguistic data science*, supported by COST (European Cooperation in Science and Technology). www.cost.eu.

8. Authors' contribution

F.A., sections 1, 2.1 (French), 2.2, 3, 4.1, 5 and 6; C.L., sections 2.1 (Hebrew) and 4.3; P.M. and B.M., sections 2.1 (Latin) and 4.2; G.V.O., sections 2.1 (Lithuanian) and 4.4; E.S.A. and C.O.T., sections 2.1 (Romanian) and 4.5; D.G., section 2.1 (Romanian). All the authors critically revised the final version of the manuscript.

9. Bibliographical References

References

Frank Abromeit, Christian Chiarcos, Christian Fäth, and Maxim Ionov. 2016. Linking the Tower of Babel: modelling a massive set of etymological dictionaries as RDF. In *Proceedings of the 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources (LDL-2016), 24 May 2016, Portorož, Slovenia*, pages 11 – 19.

Florentina Armaselu, Chaya Liebeskind, Paola Marongiu, Barbara McGillivray, Giedrė Valūnaitė Oleškevičienė, Elena Simona Apostol, and Ciprian-Octavian Truică. 2024. Linguistic Linked Open Data for Diachronic Analysis (LLODIA).

R Ashdowne. 2016. Data in online version of the 'dictionary of medieval Latin from British sources' (dmlbs).

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. (arXiv:2303.12712). ArXiv:2303.12712 [cs].

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in large language models: a comprehensive review. (arXiv:2310.14735). ArXiv:2310.14735 [cs].

Christian Chiarcos, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022a. Modelling frequency, attestation, and corpus-based information with OntoLex-FrAC. In *International Conference on Computational Linguistics*, pages 4018–4027.

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

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. (arXiv:1803.07640). ArXiv:1803.07640 [cs].

Jolanta Gelumbeckaitė, Mindaugas Šinkūnas, and Vytautas Zinkevičius. 2012. "senosios lietuvių kalbos tekstynas" (SLIEKKAS) - nauja diachroninio tekstyno samprata. *Darbai ir dienos*, 58:257–278.

Daniela Gifu. 2016a. Lexical semantics in text processing. contrastive diachronic studies on Romanian language.

Daniela Gifu. 2016b. Diachronic evaluation of newspapers language between different idioms. In *Proceedings of the IJCAI 2016 Workshop. Natural Language Processing meets Journalism*.

Anas Khan. 2018. Towards the representation of etymological data on the Semantic Web. *Information*, 9(12):304.

Anas Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena González-Blanco García, Jorge Gracia, Maxim Ionov, Penny Labropoulou, Francesco Mambrini, John P McCrae, Émilie Pagé-Perron, Marco Passarotti, Ros Salvador, and Ciprian-Octavian Truică. 2022. When linguistics meets Web technologies. Recent advances in modelling linguistic linked open data. *Semantic Web*.

Fahad Khan, Laurent Romary, Ana Salgado, Jack Bowers, Mohamed Khemakhem, and Toma Tasovac. 2020. Modelling etymology in LMF/TEI: The Grande Dicionário Houaiss da Língua Portuguesa Dictionary as a use case. In *LREC2020-12th Language Resources and Evaluation Conference*.

Charlton T. Lewis and Charles Short. 1879. *A Latin Dictionary, Founded on Andrews' edition of Freund's Latin dictionary revised, enlarged, and in great part rewritten by Charlton T. Lewis, Ph.D. and Charles Short*. Clarendon Press, Oxford.

Chaya Liebeskind, Ido Dagan, and Jonathan Schler. 2012. Statistical thesaurus construction for a morphologically rich language. In ** 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)*, pages 59–64.

Chaya Liebeskind and Shmuel Liebeskind. 2020. Deep learning for period classification of historical Hebrew texts. *Journal of Data Mining & Digital Humanities*, 2020:5864.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, page 47–58, Toronto, Canada. Association for Computational Linguistics.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, and Paul Buitelaar. 2017. The OntoLex-Lemon model: development and applications. In *Proceedings of eLex 2017 Conference*.

Barbara McGillivray and Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of Latin. In *New Methods in Historical Corpus Linguistics*, Tübingen. Narr.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv:1301.3781 [cs]*. ArXiv: 1301.3781.

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

Sabrina Ortiz. 2023. What are Microsoft's different Copilots? Here's what they are and how you can use them.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv:1802.05365 [cs]*. ArXiv: 1802.05365.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, page 45–50, Valletta, Malta. ELRA.

Thesaurusbüro München Internationale Thesaurus-Kommission, editor. 1900–. *Thesaurus linguae latinae*. Mouton de Gruyter, Berlin.

Ciprian-Octavian Truică, Victor Tudose, and Elena-Simona Apostol. 2023. Semantic change detection for the Romanian language. In *International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC2023)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. (arXiv:2201.11903). ArXiv:2201.11903 [cs].

Christopher Welty, Richard Fikes, and Selene Makarios. 2006. A reusable ontology for fluents in OWL. In *Proceedings of the Fourth International Conference, FOIS 2006*, page 8, Baltimore, Maryland, USA.