Graph Databases for Diachronic Language Data Modelling

Barbara McGillivray
King's College London, UK
barbara.mcgillivray@kcl.ac.uk

Pierluigi Cassotti and **Davide Di Pierro**
University of Bari Aldo Moro, Italy
{surname.name}@uniba.it

Fahad Khan
Istituto di Linguistica Computazionale, CNR
fahad.khan@ilc.cnr.it

Paola Marongiu
University of Neuchâtel, Switzerland
paola.marongiu@unine.ch

Stefano Ferilli and **Pierpaolo Basile**
University of Bari Aldo Moro, Italy
{surname.name}@uniba.it

Abstract

Diachronic analysis, particularly of lexical semantics, is one of the most intriguing and complex tasks in linguistic studies. The integration of lexical semantic information and diachronic language resources plays a critical role in enabling quantitative accounts of language change. Focusing on the case of Latin, a high-resource language among historical languages, we present initial results from integrating Latin corpus data, Latin WordNet, and Wikidata into a graph database via a Graph-BRAIN Schema and show the potential offered by this model for diachronic semantic research.

1 Introduction and Background

Research in empirical historical semantics requires access to various sources, from dictionaries and lexicons to encyclopedic information and diachronic texts. While several scholars have recognized the corpus-based nature of diachronic semantics, particularly for corpus languages like Latin (Pinkster, 1991; Geeraerts et al., 2012), quantitative corpus-based studies are yet to pervade historical semantics research. A critical barrier to this is that corpus and lexical resources for historical languages tend to exist in data siloes. While significant progress on linking lexical resources, tools, and corpora at the level of lemmas has been made (cf. Passarotti et al. (2020) for Latin), linking at the level of word senses is still missing.

Given the remarkable work done in the design of linked data models for language data (Khan et al., 2022), some studies such as Armaselu et al. (2022) have already advocated for integrating corpus approaches with Linked Open Data technologies to study lexical semantic change, i.e., the phenomenon concerned with the change in the meaning of words over time. One crucial strategy for representing the results of research into language change as linked data is by modeling and publishing them as knowledge bases using a lexicon-based model, usually OntoLex-Lemon and its various extensions. This includes the soon-to-be-published Frequency Attestations and Corpus (FrAC) module, which proposes a new series of classes and properties for linking elements of a lexicon with corpora (Chiarcos et al., 2022). Previous work in this area includes a proposal to modify the core organizing principles of wordnets in order to represent semantic shift phenomena (Khan et al., 2023), as well as work on the representation of etymologies as Resource Description Framework (RDF) graphs using OntoLex-Lemon (Khan, 2018) and the integration of temporal information into linguistically linked datasets via a so-called *four-dimensionalist* approach (Khan, 2020).

Integrating lexical resources and semantically-annotated corpus data at scale would allow us to gather corpus data on sense distribution information, essential for fully implementing the quantitative turn in historical semantics (McGillivray and Jenset, 2023). This integration, however, requires efficient handling of large datasets. An opportunity to combine the efficient storage, management, and retrieval of data offered by Data Base Management Systems (DBMSs) with the support for formal reasoning offered by Knowledge Bases (KBs) comes from the recent development of *Graph Databases*. Graph DBMS are intrinsically designed to store schemaless data, making them suitable to dynamic systems in which merging information is relevant. Unlike traditional DBMSs such as relational (Kriegel et al., 2003) or object-oriented (Bertino and Martino, 1991) ones, Graph DBMS lack predefined structures. Neo4j [1] is among the most common graph DBMSs. The Graph-BRAIN[2] technology (Ferilli and Redavid, 2020) provides intelligent information retrieval function-alities on a graph database. Its interface provides end users with access to data employing schema definitions. Schemes (available in terms of classes, relationships, and attributes) coordinate how data is presented in the interface. In Basile et al. (2022), we proposed the *Linguistic Knowledge Graph*, a model based on graph DBMSs. The Linguistic Knowledge Graph models relations between con-cepts and words, information about word occur-rences in corpora, and diachronic information on both concepts and words. In McGillivray et al.(2023), we show an application of this model to the lexical-semantic analysis of Latin data.

Our choice to focus on Latin is motivated by several factors. First, Latin has one of the longest recorded histories of any human language, making it naturally suitable for quantitative studies (Pinkster, 1991); this, in turn, allows for corpus-driven analyses of semantic change processes over long periods. Second, this language has a particularly favourable position among historical languages: there is a high availability of extensive Latin corpora in digital form (some of which have been linked to language resources at the level of word lemmas in the context of the LiLa project [3]) and of computational language resources such as Latin WordNet (Minozzi, 2017) and digitized dictionaries such as the Lewis & Short Latin dictionary[4].

Focusing on the development of the Latin language, in this paper we expand the range of Latin language resources included in the Linguistic Knowledge Graph for the study of lexical semantic change in Latin.[5] Our contributions include: (i) the ingestion of Latin WordNet into the Linguistic Knowledge Graph; (ii) a new curated linking between existing resources for Latin, namely Latin WordNet (Minozzi, 2017; Biagetti et al., 2021) and the SemEval 2020 Task 1 Latin dataset (McGillivray, 2021), a sense-annotated portion of the LatinISE diachronic corpus of Latin (McGillivray et al., 2022);[6] (iii) the integration of external contextual information (Wikidata) about the occupations of Latin authors. The term 'occupation' is here used in a broad sense, to refer to various types of political, cultural and societal profiles that identify authors in Wikidata. These could be e.g., priests, philosophers, historians, hagiographers, among others.

2 Resources

2.1 Dataset

LatinISE contains approximately 10 million word tokens from texts dating from the fifth century BCE to the contemporary era; it has been semi-automatically lemmatized and part-of-speech tagged. The corpus includes metadata fields indicating text identifier, author, title, dates, century, genre, URL of the source, and book title/number and character names (for plays). The semantically annotated dataset we use here was created as part of the SemEval shared task on Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020) and will be henceforth referred to as the SemEval Latin dataset. It contains in-context annotations for 40 Latin lemmas, 20 of which are known to have changed their meaning concerning Christianity (for example, *beatus*, which shifted its meaning from 'fortunate' to 'blessed'), and 20 are known not to have changed their meaning between the BCE era and the CE era. For each of these lemmas, 60 sentences were annotated, of which 30 were randomly extracted from BCE texts and 30 from CE texts. The annotation was conducted following a variation of the DuReL framework (Schlechtweg et al., 2018) described in Schlechtweg et al. (2020): the degree by which a usage instance of a target word is related to each of its possible dictionary definitions was annotated using a four-point scale (Unrelated, Distantly Related, Closely Related, and Identical). The definitions were drawn from the Logeion online dictionary (https://logeion.uchicago.edu/), which contains Lewis and Short's *Latin-English Lexicon* (1879) (Lewis and Short, 1879), Lewis' *Elementary Latin Dictionary* (1890) (Lewis, 1890), and the dictionary by Du Fresne Du Cange et al. (1883-1887). The de-

[1] https://neo4j.com/

[2] http://193.204.187.73:8088/GraphBRAIN/

[3] https://lila-erc.eu/

[4] https://lila-erc.eu/data/lexicalResources/LewisShort/Lexicon

[5] Our code and data are available at https://github.com/linguisticGraph/latin-graph

[6] Openly available at https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2506.

tails of the annotation are described in McGillivray et al. (2022).

2.2 Curated Linking

We manually linked each word sense of the SemEval Latin dataset to one or more WordNet synsets. We started with the dataset provided by the LiLa project (Franzini et al., 2019), which contains a sample of 10,314 lemmas from Latin WordNet (LWN) (Minozzi, 2017; Biagetti et al., 2021). The LiLa team verified and corrected, where necessary, the synsets associated with each lemma of the sample and linked them to version 3.0 of Princeton WordNet (PWN) (Fellbaum, 1998; Miller, 1992). However, as the LiLa dataset only covers 22 of the 40 lemmas in our dataset, we used LWN as a reference for the remaining 18 lemmas. We converted the synset codes 1.6 used by LWN to version 3.0 of PWN for consistency.

The senses assigned to the target words in the SemEval Latin dataset often condensed multiple meanings into a single definition, requiring multiple synsets to be linked to the same meaning to capture all nuances. For example, the sense "understanding, judgment, wisdom, sense, penetration, prudence" of the lemma *consilium* was linked to four synsets.

In some cases, a particular sense could not be described by any of the assigned synsets in the LiLa dataset. In such cases, we searched for the lemma in LWN and selected a more appropriate synset. This was the case e.g. for the adjective *acerbus* and one of its meanings in the SemEval Latin dataset "(of things) heavy, sad, bitter". For this meaning we selected the synset 01650376-a "psychologically painful" from LWN. When we could not find the synset in either LWN or the LiLa dataset, we looked for the most suitable synset in PWN. However, for some meanings specific to Roman culture and institutions, we could not find a suitable synset, such as with the meaning 'Virtue, personified as a deity' of *virtus*. In these cases, we did not link the sense to WordNet.

2.3 Contextual Information

In some instances, the metadata field of the SemEval Latin dataset (which indicates the author and title of the text, dating, and genre) was noisy, incorrectly structured, or incomplete. Wikidata is an extensive, collaboratively maintained knowledge base (Vrandečić and Krötzsch, 2014), hosting more than one hundred million items. We exploited Wikidata for de-noising and linking the authors of the documents containing the sentences in our dataset.

First, we extracted the Wikidata entities for which the author's occupation is specified (wdt:P106, *occupation*), and Latin (wd:Q397, *Latin*) is one of the writing languages for the author (wdt:P6886, *writing language*). We retrieve information about each author in the form of key/value properties. Author names in the SemEval Latin dataset can occur in different languages and different forms, for example *praenomen* and *nomen* followed by *cognomen* e.g., Marcus Tullius Cicero; *cognomen* followed by *praenomen* and *nomen* e.g., Cicero, Marcus Tullius; only *cognomen* e.g., Cicero; only *praenomen* and *nomen* e.g., Marcus Tullius. We processed the author's mentions in the SemEval Latin dataset and the writer labels and aliases extracted from Wikidata, performing lowercase and punctuation removal. Matching is realized by computing the Levenshtein distance (Schimke et al., 2004) between the author reported in the SemEval dataset and all the collected surface forms (i.e., labels/aliases) from Wikidata. The surface forms are then ranked by decreasing Levenshtein distance. If the Levenshtein distance between the author's mention and the top-ranked surface form is less than a fixed threshold, i.e., $\delta = 0.1$, the entity referenced by the surface form is linked to the author's mention. For each author, Wikidata provides rich information, such as biographical data, the author's works, and events that influenced their life and production. In this study, we focus on occupation information: we encode the information provided by Wikidata about the occupations of the author exploiting the property wdt:P106 (*occupation*). In particular, we create nodes of type Occupation for each occupation retrieved in Wikidata, generating a relationship between the author and their respective occupation.

3 GraphBRAIN

We stored the above information in a graph-based structure, specifically in a knowledge graph based on the GraphBRAIN technology (Ferilli and Redavid, 2020). GraphBRAIN is an approach to knowl-edge bases in graph form using a graph database (DB) to store information, coupled with an ontol-ogy that defines what information can be stored in the DB and how it must be described. Unlike the RDF graph model, traditionally used in Semantic Web approaches, GraphBRAIN adopts the Labelled Property Graph (LPG) model, where nodes and arcs may be labelled and carry information as attribute-value pairs, ensuring a more compact and human-readable representation of knowledge. The DBMS underlying GraphBRAIN is currently Neo4j (Miller, 2013), which is schema-less. GraphBRAIN proposes an XML-based formalism to express LPG ontologies that can be mapped onto the elements of LPG graphs and act as a schema for the DB (Ferilli et al., 2022b). This approach brings several advantages. The efficiency of a native LPG graph DB can be leveraged to run network analysis and graph mining algorithms. In contrast, the expressiveness of the ontology can be leveraged for advanced automated reasoning capabilities. The ontology and data can be imported from or exported to Web Owl Language (OWL), thus enabling the use of Semantic Web tools. However, they can also be imported or exported to other formalisms (e.g., Prolog), enabling different kinds of inference, e.g., rule-based deduction, abduction, abstraction, argumentation (Esposito et al., 2000).

The Linguistic Knowledge Graph (McGillivray et al., 2023) allows us to express information about corpora, linguistic properties (background lexical, morphological, syntactic, and semantic information), time, and context; linguistic information can be imported from existing resources such as WordNet. Its lexical part is inspired by and aligned to the standard ontological lexicon model OntoLex-Lemon (McCrae et al., 2014). A corpus can be described at several levels of granularity (word, sentence, text, document). Contextual information concerns the standard bibliographic metadata (e.g., authors, publishers) but may be expanded to other entities (e.g., events). Time information can describe specific time points (days, months, years, centuries) or time intervals.

3.1 Linguistic Ontology

To address the need to create a shared vocabulary to visualize and connect the data, we here describe our linguistic ontology's main components. This scheme collects all the relevant pieces of information available in standard lexical databases and other relevant sources of knowledge for diachronic analysis. We report the classes and relationships of our ontology in boldface; words are represented in lower-case, and relationships in upper-case. **Document** represents the hub for knowledge discovery since it contains most aspects of the knowledge that we need. It is linked to the **Person** who wrote the text (**HAS_AUTHOR**), commonly named the "author". A document may **CONCERN** specific **A rtifact**s, **D evice**s, belong to (**BELONGS_TO**) one **Category**, be written in at least one (**HAS_LANGUAGE**) **Language** and published (**PUBLISHED_IN**). We represent **Text**s belonging to (**BELONGS_TO**) documents. From the text, we are able to represent the **Word**s it contains. **Lemma**s are labelled with their information, e.g., morphology and **PartOfSpeech** tags. On the other hand, word forms have (**HAS_LEMMA**) lemmas. Synsets have relationships with each other; one may be a sub-synset (hyponym) of another (**IS_A**) or be equivalent to (**SAME_AS**) another one in a different database. This happens when mapping Princeton WordNet to Latin WordNet. Time needs to be modelled for diachronic analysis. **TemporalSpecification** includes **TimeIntervals** and specific **T imePoint**s, n amely **Year**, **Month**, and **Day**. This model allows authors and texts to be bound to specific time p eriods. Moreover, we have **Event**s, which may come in handy to understand the reason why some words changed their meaning (e.g., in relation to Christianity).

3.2 Latin WordNet Ingestion

The Latin WordNet (LWN) project is an initiative to create and share a common lexico-semantic database of the Latin language. The project originated as a branch of the MultiWordNet (Pianta et al., 2002) project. For diachronic analyses, linking linguistic resources with temporal information allows us to uncover instances of semantic changes in the usage of words. Hence, we provide a mechanism to enrich the Linguistic Knowledge Graph with Latin WordNet and exploit the hierarchical structure of the relationships between synsets.

In Section 3, we described the GraphBRAIN tech-nology and its reliance on schemes/ ontologies to deliver information extraction and reasoning functionalities. We mapped the Latin WordNet data with the portion of our ontology specifi-cally devoted to linguistic analysis and understand-ing. Further details about scheme specifications for document representation are available in (Fer-illi et al., 2022a). Here we describe the map-ping between the lexical database and our schema. In LWN, we identified the following resources, grouped into separate Comma Separated Value

(CSV) files: lemma, lexical_relation, literal_sense, metaphoric_sense, metonymic_sense, phrase, semantic_relation, synset. Each resource has features that may be seen as classical columns in a relational database. From now on, we refer to specific fields as *resource.field* to uniquely identify them and motivate how we map them. The alignment process is as follows:

- lemma: a specific lemma is embedded in our class **Lemma**. A **Lemma** is characterized by a unique id, a lemma (its value), and a PoS tag (modelled as a relationship). For our purposes, the class **PartOfSpeech** collects all the pos tags used, following the Universal PoS Tags standard[7]. We can represent other fields expressed in LWN, such as *lemma.uri*.

- lexical_relation: this represents a relationship between two **Lemma**s. The field *lexical_relation.type* specifies the type of relationship. We modelled the present ones with some explicit names which express their meanings: **ANTONYMOUS_OF**, **PERTAINS_TO** (to refer to the type of relation indicated by the attribute of the relations), with their corresponding inverses, e.g. **IS_PAST_PARTICIPLE_OF**.

- literal_sense: this represents a relationship between a lemma, identified by the field *literal_sense.lemma*, and a synset, identified by *literal_sense.synset*. We call this relationship **expresses**. We highlight that the relationship has a "literal" sense by adding a specific attribute **sense**. Additional information about the period and genre is available.

- metaphoric_sense: similarly to the previous one, this represents a relationship between a lemma and a synset, where the **sense** is "metaphoric".

- metonymic_sense: as before, but the **sense** is "metonymic" in this case.

- phrase: a phrase is a word or a multi-word expression. In both cases, the concept is expressed by the class **Lemma** since for our purposes both concepts play an equally important role when analysing semantic changes. Again, we have the PoS tag information, which is modelled in the same way described above.

- semantic_relation: a relationship between two synsets. Based on the *semantic_relation.type* several relationships may be expressed. They are mapped into the following ones and their corresponding inverses: **PART_OF**, **HAS_SUBCLASS**, **ATTRIBUTE_OF**, **SIMILAR_TO**, **ANTONYMOUS_OF**, **PERTAINS_TO**, **PART_PARTICIPLE_OF**, **CAUSES**, and **ENTAILS**.

- synset: a synset is embedded in **LexiconConcept** while its property synset.gloss, which is the description of the synset, is represented as the attribute **description** of the class **LexiconConcept**. *synset.gloss* is the description of the synset and is mapped onto the attribute **description**.

Thanks to this mapping, we can acquire the LWN resource and represent it in our formalism, which allows us to leverage the connections between the different datasets, as explained via examples in the next section.

4 Analysis and Discussion

Figure 1 shows the subgraph for the word *humanitas*. The occurrences of *humanitas* are annotated in the SemEval dataset with three senses: (i) 'human nature, humanity', (ii) 'humanity, philanthropy', and (iii) 'mankind'.[8] In the curated link, we associate the sense (i) to the humanness.n.01 synset, the sense (ii) to the synsets kindness.n.01, kindness.n.03, and courtesy.n.03 and sense (iii) to the synset world.n.08. According to the *Thesaurus Linguae Latinae* (Thesaurus-Kommission, 1900–), which confirms the first attestation of all senses in the 1st century BCE, the sense (ii) 'humanity, philanthropy' developed from the more general sense (i) 'human nature, humanity' which refers to human nature in general. The subgraph shows that the three senses are attested at least once in passages dated 1st century BCE. However, the graph shows that the sense of 'philanthropy' dominates all other senses in the 1st century BCE. In the transition to the CE period, the sense of 'humanity' prevails

[7] https://universaldependencies.org/u/pos/

[8] A fourth sense 'liberal education, good breeding, the elegance of manners or language, refinement' was annotated in the Latin dataset, but not encoded in the graph, since the author matching described in Section 2.3 failed.

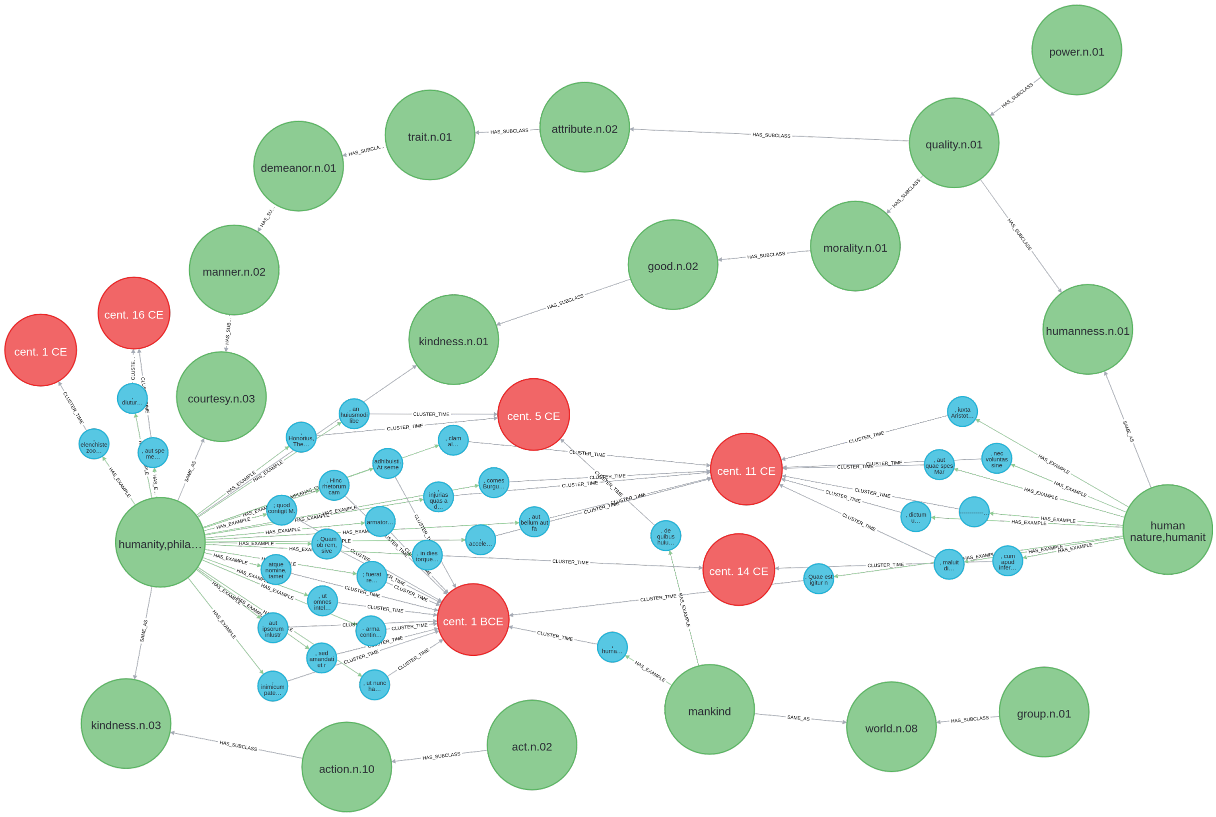
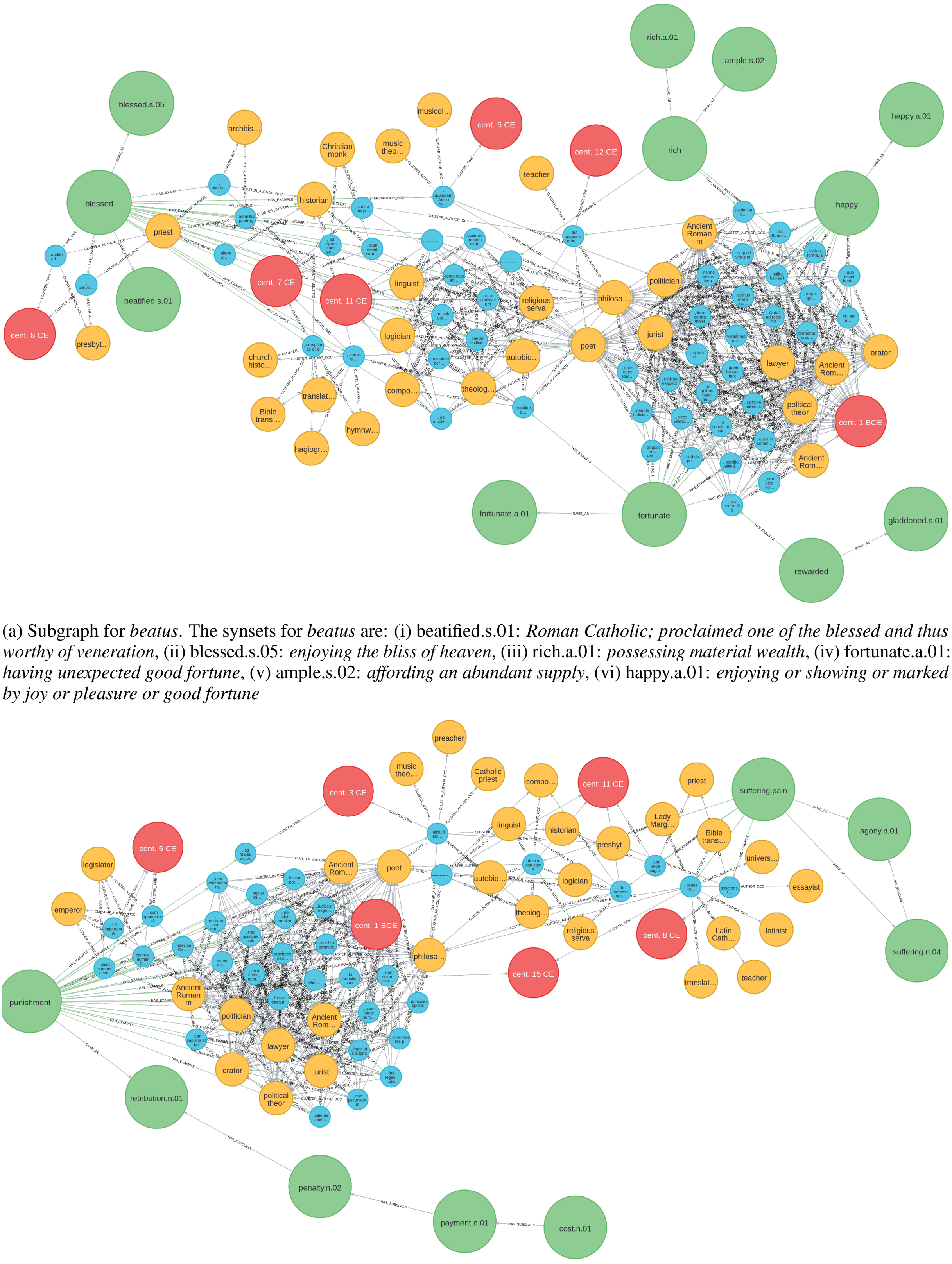


Figure 1: Subgraph for the word *humanitas*, including the sentences in which the lemma *humanitas* occurs in the SemEval Latin dataset, the century of the works from which the sentences were extracted, the annotated senses in the SemEval Latin dataset, and the curated links between the senses and the synsets in Latin WordNet. The sentences are represented as Text nodes (in blue), the senses and the synsets as LexiconConcept nodes (in green), and the centuries as TimePoint nodes (in red).

regarding the number of annotations, and the two meanings coexist in the CE period.

By ascending the WordNet hierarchy, we can gain deeper insight into the relationship between the two senses. The sense (ii) ‘humanity, philanthropy’ and the sense (i) ‘human nature’ are connected via two paths: sense (ii) originates from the quality.n.01 synset (i.e. ‘an essential and distinguishing attribute of something or someone’); sense (i) from the attribute.n.02 synset (i.e., ‘an abstraction belonging to or characteristic of an entity’). The two senses have in common the quality.n.01 synset, but the sense (ii) ‘humanity, philanthropy’ is directly linked to kindness.n.01 synset, and to a higher degree of the WordNet hierarchy to the morality.n.01 synset (i.e., ‘concerned with the distinction between good and evil or right and wrong’). The additional information provided by including the WordNet hierarchy in the graph allows us to show the type of semantic relationship between the two predominant senses of *humanitas*. The more general sense (i) ‘human nature’ specializes in its meaning in the sphere of morality, originating the sense (ii) ‘philanthropy’. In the example of *humanitas* shown in Figure 1, the injected information from WordNet was exploited to analyze the semantic relationship between the meanings of the lemma *humanitas*. While the synset taxonomy in this example helps us track and classify phenomena of semantic change, including other types of information retrievable from the metadata can help gain further insights into the context of the semantic change. We add information about the authors’ occupations in the examples shown in Figure 2.

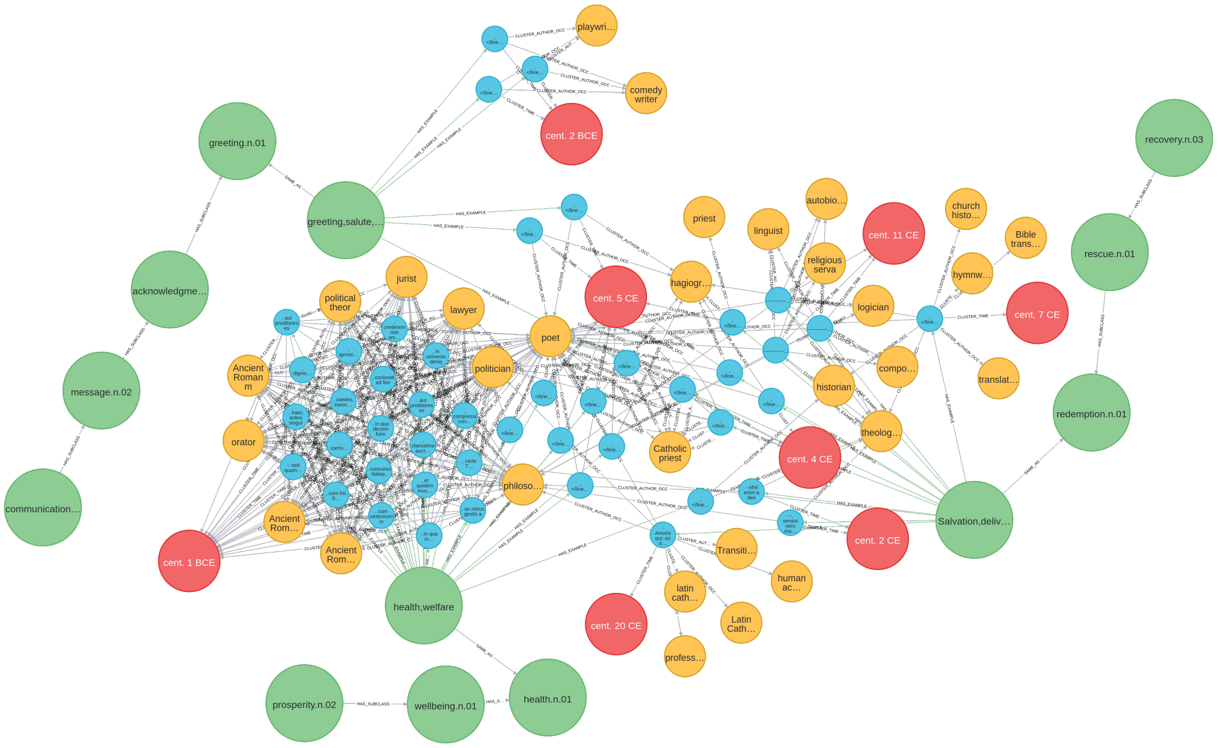
In Figure 2, three examples of subgraphs are shown. The three graphs refer to the encoded information for the Latin lemmas *beatus*, *poena*, and *salus,* respectively. In particular, we filtered for nodes of type Text (blue nodes), Century (red nodes), Synset (green nodes), and Occupation (yellow nodes). We grouped the Text nodes by occu-pation and century, i.e., we created an explicit link between nodes of type Text and nodes of type Time-Point and between nodes of type Text and nodes of



(a) Subgraph for *beatus*. The synsets for *beatus* are: (i) beatified.s.01: *Roman Catholic; proclaimed one of the blessed and thus worthy of veneration*, (ii) blessed.s.05: *enjoying the bliss of heaven*, (iii) rich.a.01: *possessing material wealth*, (iv) fortunate.a.01: *having unexpected good fortune*, (v) ample.s.02: *affording an abundant supply*, (vi) happy.a.01: *enjoying or showing or marked by joy or pleasure or good fortune*

(b) Subgraph for *poena*. The synsets for *poena* are: (i) retribution.n.01: *a justly deserved penalty*, (ii) suffering.n.04: *feelings of mental or physical pain*, (iii) agony.n.01: *intense feelings of suffering; acute mental or physical pain*

Figure 2: Sub-graphs: (a) beatus. (b) poena (c) salus .



(c) Subgraph for *salus*. The synsets for *salus* are: (i) health.n.01: *a healthy state of well-being*, (ii) redemption.n.01: *(Christianity) the act of delivering from sin or saving from evil*, (iii) greeting.n.01: *an acknowledgment or expression of goodwill*

Figure 2: Sub-graphs: (a) beatus. (b) poena (c) salus (cont.).

type Occupation.

Combining queries at the level of the annotated senses, WordNet synsets, text metadata and textual data at once, users can have access to rich nuanced information, which is very valuable for quantitative diachronic semantic analyses, both on specific words and whole lexical fields. The graphs in Figure 2 seem to show some trends in semantic change, all related to Christianity. The lemma *beatus* was annotated in the SemEval dataset with five senses: (i) 'happy,' (ii) 'fortunate', (iii) 'rewarded', (iv) 'rich', and (v) 'blessed'. The graph shows that the senses (i) 'happy', (ii) 'fortunate', (iii) 'rewarded', and (iv) 'rich' all emerge starting from the 1st century BCE in the annotated dataset. On the other hand, sense (v) 'blessed' emerges later with the advent of Christianity, as we can see in correspondence with the CE nodes. In this case, there seems to be a replacement of the previous senses in favour of the Christian sense. Additionally, if we consider the nodes of type Occupation, a noticeable difference emerges between the two (groups of) meanings: in the cluster of occupation nodes connected to the Christian sense, we can observe profiles related to theological and religious activity, e.g., priests, hagiographers, which do not appear to be connected to the other senses. The same type of observations can be made for *salus*, which initially has the meanings (i) 'health' and (ii) 'greeting', and, subsequently, develop the Christian sense of (iii) 'salvation, deliverance from sins'. However, in this case, we can notice the difference with *beatus* in the type of semantic change, as the new meaning (iii) 'salvation' replaces or dominates the previously attested meanings but continues to coexist with them. The lemma *poena* also presents an example of semantic change in which the new meaning does not entirely replace the previous ones. The new sense of 'suffering, pain', which emerges in the CE nodes, continues to coexist with the sense of 'punishment', which was attested from the 1st century BCE in the annotated dataset. In the case of *poena*, the contrast between the two clusters of occupation nodes is even more evident. The sense of punishment is often associated with authors classified as related to the legal world, e.g., legislator, lawyer, and jurist. In contrast, nodes related to the Christian and theological world appear in the case of salvation, e.g., theologian, priest, and presbyter. The graphs in Figure 2 are in line with that we know about semantic changes prompted by the advent of Christianity, which invested many words already in use in pre-Christian Latin with new meanings closely related to the Christian world (Burton, 2011). Moreover, the lemmas shown in Figure 2 illustrate the different types of interaction between older and new senses described in literature (Traugott and Dasher, 2001, 10–12): in some cases, the two senses can continue to coexist, as for the lemmas *salus* and *poena* (a phenomenon called 'layering' (Hopper, 1991, 22)); in others, as for the lemma *beatus*, the relationship between the new sense and the older ones is unbalanced as the new sense becomes more prominent in a society invested in Christian values.

5 Conclusion and Future Work

We applied diachronic lexical-semantic analysis by integrating different resources into a graph-based structure. Future research should be devoted to enriching the dataset by collecting other resources to uncover more complex relationships and possibly automatically detect semantic changes among all terms in the vocabulary. Currently, our model does not include a programmatic way to automatically detect instances of semantic changes, but this is an avenue of future research. We plan to publish a version of the graph database in which experiments can be replicated.

Authors' contributions and Acknowledgements

BMcG contributed to the design of the study, managed the project, provided the SemEval dataset and wrote sections 1 and 2.1. PM provided the curated linking between the annotated SemEval Latin dataset and WN, and wrote sections 2.2 and 4. PC processed the annotated LatinISE corpus, extracted metadata information from WikiData, generated the graph and the visualizations (Figure 1 and Figure 2), wrote Section 2.3, and contributed in writing Section 4. PB contributed to the design of the study, generated the graph and wrote section 3. FK proofread the article and contributed to discussions on the relationship between native KG approaches to modelling lexical data as graphs and RDF/OntoLex approaches. DD contributed to the design of the schema and the upload of LWN resources into the LPG-based KG and wrote section 3. SF contributed to the design of the schema and the Knowledge Representation methodology of GraphBRAIN and wrote section 3.

We acknowledge the support of the PNRR projects FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) and CHANGES - Cultural Heritage Active innovation for Next-GEn Sustainable society (PE00000020), Spoke 3 - Digital Libraries, Archives and Philology, under the NRRP MUR program funded by the NextGenerationEU.

References

Florentina Armaselu, Elena Simona Apostol, Anas Fahad Khan, Chaya Liebeskind, Barbara McGillivray, Ciprian-Octavian Truica, Andrius Utka, Giedre Valunaite Oleskeviciene, and Marieke van Erp. 2022. LL(O)D and NLP perspectives on semantic change for humanities research. *Semantic Web*, 13(6):1051–1080.

Pierpaolo Basile, Pierluigi Cassotti, Stefano Ferilli, and Barbara McGillivray. 2022. A new time-sensitive model of linguistic knowledge for graph databases. In *Proceedings of the 1st Workshop on Artificial Intelligence for Cultural Heritage, AI4CH 2022, co-located with the 21st International Conference of the Italian Association for Artificial Intelligence (AIxIA 2022), Udine, Italy, November 28, 2022*, volume 3286 of *CEUR Workshop Proceedings*, pages 69–80. CEUR-WS.org.

Elisa Bertino and Lorenzo Martino. 1991. Object-oriented database management systems: concepts and issues. *Computer*, 24(4):33–47.

Erica Biagetti, Chiara Zanchi, and William Michael Short. 2021. Toward the creation of WordNets for ancient Indo-European languages. In *Proceedings of the 11th Global Wordnet Conference*, pages 258–266, University of South Africa (UNISA). Global Wordnet Association.

Philip Burton. 2011. Christian latin. In *A companion to the Latin language*, pages 485–501, Oxford. Wiley-Blackwell.

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

Charles Du Fresne Du Cange, G. A. Louis Henschel, P. Carpentier, Johann Christoph Adelung, and Léopold Favre. 1883-1887. *Glossarium mediæet infimælatinitatis*. L. Favre, Niort.

Floriana Esposito, Giovanni Semeraro, Nicola Fanizzi, and Stefano Ferilli. 2000. Multistrategy theory revision: Induction and abduction in INTHELEX. *Mach. Learn.*, 38(1-2):133–156.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Stefano Ferilli and Domenico Redavid. 2020. The graphbrain system for knowledge graph management and advanced fruition. In *Foundations of Intelligent Systems: 25th International Symposium, ISMIS 2020, Graz, Austria, September 23–25, 2020, Proceedings*, pages 308–317. Springer.

Stefano Ferilli, Domenico Redavid, and Davide Di Pierro. 2022a. Holistic graph-based document representation and management for open science. *International Journal on Digital Libraries*, pages 1–23.

Stefano Ferilli, Domenico Redavid, and Davide Di Pierro. 2022b. Lpg-based ontologies as schemas for graph dbs. In *Proceedings of the 30th Italian Symposium on Advanced Database Systems, SEBD 2022, Tirrenia (PI), Italy, June 19-22, 2022*, volume 3194 of *CEUR Workshop Proceedings*, pages 256–267. CEUR-WS.org.

Greta Franzini, Andrea Peverelli, Paolo Ruffolo, Marco Passarotti, Helena Sanna, Edoardo Signoroni, Viviana Ventura, and Federica Zampedri. 2019. *Nunc Est Aestimandum*: Towards an evaluation of the latin wordnet. In *Proceedings of the Sixth Italian Conference on Computational Linguistics*. Accademia University Press.

Dirk Geeraerts, Caroline Gevaert, and Dirk Speelman. 2012. Current methods in historical semantics. *Current methods in historical semantics*, pages 73–109.

Paul J. Hopper. 1991. On some principles of grammaticalization. In *Approaches to grammaticalization*, pages 17–35, Amsterdam, Philadelphia. John Benjamins Publishing.

Anas Fahad Khan. 2018. Towards the representation of etymological data on the semantic web. *Information*, 9(12):304. Publisher: MDPI AG.

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, Salvador Rosl Muñoz, and Ciprian-Octavian Truică. 2022. When linguistics meets web technologies. recent advances in modelling linguistic linked data. *Semantic Web*, pages 1–64.

Anas Fahad Khan, John P McCrae, Francisco Javier Minaya Gómez, Rafael Cruz González, and Javier E Díaz-Vera. 2023. Some considerations in the construction of a historical language wordnet.

Fahad Khan. 2020. Representing temporal information in lexical linked data resources. In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 15–22, Marseille, France. European Language Resources Association.

Hans-Peter Kriegel, Martin Pfeifle, Marco Pötke, and Thomas Seidl. 2003. The paradigm of relational indexing: A survey. In *BTW 2003–Datenbanksysteme für Business, Technologie und Web, Tagungsband der 10. BTW Konferenz*. Gesellschaft für Informatik eV.

Charlton T. Lewis. 1890. *An Elementary Latin Dictionary*. American Book Company, New York, Cincinnati, and Chicago.

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.

John McCrae, Christiane Fellbaum, and Philipp Cimiano. 2014. Publishing and linking wordnet using lemon and rdf. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*.

Barbara McGillivray. 2021. Dataset: Latin lexical semantic annotation. Figshare. DOI: https://doi.org/10.18742/16974823.v1.

Barbara McGillivray, Pierluigi Cassotti, Pierpaolo Basile, Davide Di Pierro, and Stefano Ferilli (in press). 2023. Using graph databases for historical language data: Challenges and opportunities. In *Proceedings of the 19th Italian Research Conference on Digital Libraries, Bari, Italy, February 23-24, 2023*, CEUR Workshop Proceedings. CEUR-WS.org.

Barbara McGillivray and Gard B. Jenset. 2023. Quantifying the quantitative (re-)turn in historical linguistics. *Humanities and Social Sciences Communications*, 10(37).

Barbara McGillivray, Daria Kondakova, Annie Burman, Francesca Dell'Oro, Helena Bermúdez Sabel, Paola Marongiu, and Manuel Márquez Cruz. 2022. A new corpus annotation framework for latin diachronic lexical semantics. *Journal of Latin Linguistics*, 21(1):47–105.

George A. Miller. 1992. WORDNET: a lexical database for english. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, USA, February 23-26, 1992*. Morgan Kaufmann.

Justin J Miller. 2013. Graph database applications and concepts with neo4j. In *Proceedings of the southern association for information systems conference, Atlanta, GA, USA*, volume 2324.

Stefano Minozzi. 2017. Latin wordnet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'information retrieval. In *Strumenti digitali e collaborativi per le Scienze dell'Antichita*, pages 123–134, Venezia. Università Ca' Foscari.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin. *Studi e Saggi Linguistici*, 58.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Harm Pinkster. 1991. *Sintassi e semantica latina*. Rosenberg & Sellier.

Sascha Schimke, Claus Vielhauer, and Jana Dittmann. 2004. Using adapted levenshtein distance for online signature authentication. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 931–934. IEEE.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 1–23. International Committee for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 169–174, New Orleans, Louisiana.

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

Elizabeth Closs Traugott and Richard B. Dasher. 2001. *Regularity in semantic change*. Cambridge University Press, Cambridge.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85. Publisher: ACM New York, NY, USA.