Using Graph Databases for Historical Language Data: Challenges and Opportunities

Barbara McGillivray¹, Pierluigi Cassotti²*, Pierpaolo Basile², Davide Di Pierro² and Stefano Ferilli²

¹King’s College London, Strand Campus, Strand, London, WC2R 2LS, United Kingdom
²University of Bari Aldo Moro, Department of Computer Science, Via E. Orabona 4, Bari, 70125, Italy

Abstract
The integration of semantic information into language resources has the potential to open up new avenues of enquiry into the mechanisms of language change. We present the first experiments in integrating data from Latin textual corpora and language resources into a graph database via the GraphBRAIN Schema and show the potential of this model for research into the mechanisms of semantic change in Latin.

Keywords
Knowledge Graphs, Latin Corpora, Semantic Change Detection, Graph Data Model

1. Introduction

Research in Historical Linguistics often requires the analysis and support of heterogeneous data and tools, such as lexical resources, encyclopaedias, and large corpora. Nevertheless, these resources are often siloed. Graph Databases present an ideal opportunity to combine the advantages of DataBase Management Systems (DBMSs) for handling individuals (scalability, storage optimization, efficient handling, mining and browsing of the data, etc.) with the high-level functionalities available in Knowledge Bases (KBs). Graph DBMS are intrinsically designed to store schemaless data. Differently from traditional DBMSs like the relational [1] or object-oriented [2] ones, they lack predefined structures. Following this approach, Neo4j ¹, one of the most common graph DBMSs, does not provide support for introducing ontology definitions based on labels and/or arcs. The absence of a schema may lead to ambiguity when reading and managing data in downstream applications due to the inherent ambiguity of the words used for expressing concepts. Hence, the semantics becomes blurred.

To address these issues, we propose the use of GraphBRAIN [3] as a solution. GraphBRAIN consists of a graph database which follows the Labelled Property Graph (LPG) [4] structure.

¹https://neo4j.com/
This structure stores nodes with specific labels, arcs which represent relationships among nodes and properties on both nodes and arcs. Properties are stored in the format of key/value pairs. GraphBRAIN requires KB designers to define a data schema which operates also as an ontology. GraphBRAIN provides a mapping mechanism for exporting schemes into an SW-compliant language, the Web Ontology Language (OWL). These schemes guide access through all the CRUD operations on the database but also ensure interpretability and interoperability among different applications. Following the schemes, applications become compliant with each other. Neo4j, in its Enterprise Edition, does not provide any constraint definition process. In other versions, it supports a few constraints like unique node property constraints, node property existence constraints, relationship property existence constraints and node key constraints. Evidently, these tools are not as expressive as ontology definitions.

In [5], we adopted GraphBRAIN technology to model time-sensitive linguistic knowledge in a graph database, describing a time-sensitive model of linguistic knowledge that can be used for graph databases. In this paper, we show an application of this model to the lexical semantic analysis of Latin data, i.e. the analysis of the meanings of Latin words. Differently from previous approaches, such as Basile et al. [6], Hamilton et al. [7], and Carlo et al. [8], we exploit graph database potentialities to detect semantic changes in specific concepts.

Latin is in a particularly favourable position among historical languages for the large-scale analysis of semantic change processes, thanks to a number of factors. First, Latin researchers now enjoy unprecedented access to digital data covering over two thousand years of history. Thanks to the ERC-funded LiLa project 2, seven Latin language resources and six corpora have been linked at the level of word lemmas so far, making Latin a unique case among historical languages. Second, we have access to extensive computational language resources for Latin, Latin WordNet [9], and digitised dictionaries of Latin, which provide rich information about words’ semantics and examples of usage. Finally, focussing on Latin allows us to investigate semantic change processes over long time spans. Latin has one of the longest recorded histories of any human language, making it naturally suitable for quantitative studies [10]. The first inscripational records date from the sixth century BCE, and Latin continues to be used to the current day by the Catholic Church and some academic and legal institutions around the world. Written Latin diverged from the spoken vernaculars in the second half of the first millennium of the Christian era, but it remained in use as one of the principal channels of communication across most of Europe for the next thousand years. The humanists’ conscious effort to reproduce Classical Latin led to a range of interesting developments, particularly affecting the neo-Latin lexicon to enable the expression of new concepts [11]. This extensive chronological span has raised the question of the extent to which Latin is seen as a dead or fossilised language (e.g. Herman [12], Butterfield [13]). However, it remains an open question to what extent this fossilisation affected the semantics of words, as we know that the Latin lexicon, in this respect, has remained dynamic (over 4,500 words have acquired new meanings since the Renaissance; Demo 2022). The extent to which post-classical Latin can really be considered as a “fixed” language (Leonhardt [14], Roelli [15], Langslow [16]) from the point of view of its ability to generate new meanings of words is still largely unknown beyond anecdotal evidence.

In Section 2 we present the Linguistic Knowledge Graph, in Section 3 we describe the Latin

2https://lila-erc.eu/
data that we worked on, and in Section 4 we show how we loaded the Latin data into the Linguistic Knowledge Graph. Finally, in Section 5 we draw some conclusions and outline future directions of work.

2. The Linguistic Knowledge Graph

The Linguistic Knowledge Graph (LKG) aims to capture different aspects of lexical resources, such as relations between words and concepts, morphological, and syntactical information. Moreover, LKG covers diachronic aspects of language, such as the date of publication of a document, and the birth and death of an author. The schema we designed takes inspiration from the ontological lexicon model LeMON [17]. For space constraints, we report in Table 1 node types and in Table 2 the relationships adopted for diachronic analysis. The lexical unit is represented as node of type InflectedWord or Lemma, which are subclass of Word, i.e. Lemma IS_A Word and InflectedWord IS_A Word. The Lemma can be a multi-word expression (mwe), in this case, the flag mwe is set to True. The respective lemma of an InflectedWord can be retrieved exploiting the relationship HAS_LEMMA between InflectedWord and Lemma. The LexiconConcept is used to represent the word’s meanings, and each instance of LexiconConcept represents a different meaning. For example, the LexiconConcept can represent the senses reported on a sense inventory, e.g. synsets in WordNet [18]. The relationship between a word and its meaning is expressed using the relationship HAS_CONCEPT among instances of Word and instance of LexiconConcept. Multiple relationships can be defined over couples of LexiconConcept using the reflexive relationship SEM_RELATION. At the same time, reflexive relationships over the Word instances can be described by the LEX_RELATION relationship.

The document structure from which words are extracted can be represented at different levels of granularity: Sentence, Text, Document, and Corpus. In particular, each excerpt can be represented as Text or Sentence, which is a subclass of Text. A Text may belong to (BELONG_TO) a Document and a Document can be part of (BELONG_TO) a Corpus. The occurrences of a word in a particular Text are traced by the relationship HAS_OCCURRENCE among Word and Text. In the case of sense-annotated corpora, such as SemCor, is possible to specify the occurrences of senses using the relationship HAS EXAMPLE among LexiconConcept and Text. Currently, the LKG takes into account two types of metadata: author and language. The relationship HAS AUTHOR among nodes of type Text and nodes of type Person determines the author of a Text. The relationship HAS LANGUAGE among nodes of type Text, Document, Corpus, and Word to nodes of type Language specifies the respective language.

The time is modelled using two classes of nodes: TimeInterval, and TimePoint, both subclasses of TemporalSpecification. The TimeInterval type is used when the date is not precisely stated, while the TimePoint is used in cases where the date is fixed. The start and end extremes of the TimeInterval nodes can be specified using the respective relationships startTime and endTime. In the current version of the LKG, time specification is supported for Person and Text. More specifically, the date of birth and death of authors is specified using the relationship BORN and DIED between Person and TemporalSpecification. The publishing date of a text is specified by the relationship PUBLISHED_IN among Text nodes and TemporalSpecification nodes.
3. Latin data

The data we loaded into the graph consists of a portion of the LatinISE corpus [19] annotated at the level of dictionary senses. LatinISE is a Latin corpus covering the period from the fifth century BCE to the twenty-first century and contains 10 million word tokens, semi-automatically lemmatised and part-of-speech tagged. The metadata fields in LatinISE indicate text identifier, author, title, dates, century, genre, url of source, and optionally book title/number and character names (for plays). The annotated dataset was produced as part of the SemEval shared task on Unsupervised Lexical Semantic Change Detection [20]. 40 Latin lemmas (“target words”) are selected, of which 20 are known to have changed their meaning with the advent of Christianity (for example, *beatus*, which shifted its meaning from ‘fortunate’ to ‘blessed’) and 20 are known to not have changed their meaning between the BCE era and the CE era. For each of the 40 lemmas, 60 sentences are randomly extracted from LatinISE, 30 of them are from texts dated in the BCE era, and 30 from texts dated in the CE era. Each sentence was annotated by at least one expert annotator, according to the DuReL framework [21]. The annotators were asked to judge the semantic relatedness of an instance of usage of a target word with respect to the list of its dictionary definitions using a four-point scale (Unrelated, Distantly Related, Closely Related, and Identical). The definitions were taken from the Latin portion of the Logeion online dictionary (https://logeion.uchicago.edu/) containing Lewis and Short’s *Latin-English Lexicon* (1879) [22], Lewis’ *Elementary Latin Dictionary* (1890) [23], and Du Fresne Du Cange et al. [24]. See McGillivray et al. [25] for further details about the dataset and its annotation framework.
Table 2
LKG relationships with their respective subject, object and attributes.

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Subject</th>
<th>Object</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>IS_A</td>
<td>Sentence</td>
<td>Text</td>
<td>id:Integer</td>
</tr>
<tr>
<td></td>
<td>Lemma</td>
<td>Word</td>
<td>id:Integer</td>
</tr>
<tr>
<td>BELONG_TO</td>
<td>Text</td>
<td>Document</td>
<td>id:Integer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Corpus</td>
<td>id:Integer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Category</td>
<td></td>
</tr>
<tr>
<td>HAS_OCCURRENCE</td>
<td>Word</td>
<td>Text</td>
<td>begin:Integer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>end:Integer</td>
</tr>
<tr>
<td>[LEX_RELATION]</td>
<td>Word</td>
<td>Word</td>
<td></td>
</tr>
<tr>
<td>HAS_LEMMA</td>
<td>Word</td>
<td>Lemma</td>
<td></td>
</tr>
<tr>
<td>HAS_CONCEPT</td>
<td>Word</td>
<td>LexiconConcept</td>
<td>grade:Float</td>
</tr>
<tr>
<td>HAS_EXAMPLE</td>
<td>LexiconConcept</td>
<td>Text</td>
<td></td>
</tr>
<tr>
<td>REFER_TO</td>
<td>LexiconConcept</td>
<td>Concept</td>
<td></td>
</tr>
<tr>
<td>[SEM_RELATION]</td>
<td>LexiconConcept</td>
<td>LexiconConcept</td>
<td></td>
</tr>
<tr>
<td>PUBLISHED_IN</td>
<td>Text</td>
<td>Document</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Corpus</td>
<td></td>
</tr>
<tr>
<td>HAS_LANGUAGE</td>
<td>Text</td>
<td>Document</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Corpus</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Word</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Language</td>
<td></td>
</tr>
</tbody>
</table>

MATCH
(centuryNode:TimeInterval)-[:startTime]->(startCentury:TimePoint),
(centuryNode:TimeInterval)-[:endTime]->(endCentury:TimePoint),
(pubNode:TimeInterval)-[:startTime]->(startPub:TimePoint),
(pubNode:TimeInterval)-[:endTime]->(endPub:TimePoint),
(text:Text)-[:PUBLISHED_IN]->(pubNode)
WHERE
centuryNode.description="century"
WITH text,
  centuryNode,
CASE WHEN endPub.Year > endCentury.Year THEN endCentury.Year ELSE endPub.Year END as minEnd,
CASE WHEN startPub.Year > startCentury.Year THEN startPub.Year ELSE startCentury.Year END as maxStart
WITH *
CASE WHEN minEnd-maxStart+1 > 0 THEN minEnd-maxStart+1 ELSE 0 END as time_overlap
ORDER BY time_overlap DESC
WITH text,
collect([{century:centuryNode}][0])[0] AS max
WITH *
  max.century AS century
CREATE (text)-[r:CLUSTER]->(century)
RETURN text, century
UNION ALL
MATCH
(centuryNode:TimeInterval)-[:startTime]->(startCentury:TimePoint),
(centuryNode:TimeInterval)-[:endTime]->(endCentury:TimePoint),
(text:Text)-[:PUBLISHED_IN]->(point:TimePoint)
WHERE
centuryNode.description="century" and
point.Year=startCentury.Year and
point.Year=endCentury.Year
WITH text, centuryNode as century
CREATE (text)-[r:CLUSTER]->(century)
RETURN text, century;

Listing 1: Clustering publishing date by centuries
4. Loading the Latin data in the Linguistic Knowledge Graph

For each instance of the target words in the Latin corpus we encode:

- the author as Person,
- the manuscript as Document,
- the year as TimePoint if the date is certain, TimeInterval otherwise,
- the sentence (left context, target word and right context) as Text,
- the definitions of the Lewis and Short Dictionary as LexiconConcept,
- the word lemma as Lemma,
- the inflected forms of the target words as InflectedWord,
- the scores associated with each LexiconConcept as properties of the HAS_EXAMPLE and HAS_OCCURRENCE relationships.

In order to simplify and make the visualisation more effective, we created the HAS_EXAMPLE relationship only in cases where the annotation reported a score of 4. In addition, to make more evident the distribution of senses with respect to centuries, we associate each date of publication of the texts with the reference century. We do this via the query given in Listing 1. In case a Text is not associated with a specific TimePoint, it will be linked with the century having the greatest overlap with the TimeInterval of the text itself. On the other hand, for texts for which a precise date is specified, the query associates the Text with the respective century of
its year. The centuries are represented as `TimeInterval`, and the description attribute is validated with "century". A new relationship, called `CLUSTER`, is so created among nodes of type `Text` and nodes of type `TimeInterval` to indicate the century.

A subgraph for the word `beatus` is shown in Figure 1. The graph shows the nodes representing the texts from which the word `beatus` is extracted, the centuries and the senses given in the Lewis and Short Dictionary. The relationships among these nodes are `CLUSTER` and `HAS_EXAMPLE`. The former connects nodes of type `TimeInterval` and nodes of type `Text`, see 1. The latter links `LexiconConcepts` and `Texts`. Most occurrences of the word `beatus` in the reference corpus are dated 1st century BCE and 11th century CE. One can immediately notice a difference in the distribution of the senses: “happy” and “fortunate” on the one hand are associated with the time period BCE (see the cluster of nodes on the left of Figure 1), and “blessed”, on the other hand, is associated with the time period CE (see the cluster of nodes on the right of Figure 1). In fact, only one sentence in the dataset displays the sense “blessed” in the first century BCE. Similarly, only two sentences dated CE contain the word `beatus` with the meaning of “fortunate”, the latter, on the other hand, is dated 1079-1142 CE and is an excerpt from the Sermones of Petrus Abaelardus.

5. Conclusions

In this work, we introduced an application of LKG for Latin data. It appears to be an interesting and novel approach to tackling the analysis of diachronic corpora. Furthermore, differently from previous approaches, it gives rise to explainable results since we take advantage of explicit relationships modelled as graphs. The LKG seems to lead to promising results, and it is ready for further investigations into Lexical Semantic Change Detection (LSCD). Future developments include a better visualization of resources, machine-learning-based techniques for automatic LSCD and an interface for querying and analysing the LKG data.

Acknowledgement

This work fulfils the research objectives of the PNRR project FAIR - Future AI Research, spoke 6 - Symbiotic AI, CUP H97G22000210007, as well as the CHANGES - Cultural Heritage Active Innovation for Next-Gen Sustainable Society, CUP H53C22000860006.

References


