Enhancing Linguistic Resources for Diachronic Analysis via Linked Data

Eleonora Ghizzota and Pierpaolo Basile and Claudia d'Amato and Nicola Fanizzi
Department of Computer Science, University of Bari Aldo Moro, Italy
e.ghizzota@phd.uniba.it, {name.surname}@uniba.it

Abstract

The Linked Linguistic Knowledge Graph (LLKG) is a language-independent linguistic resource designed for diachronic analysis set on well-known ontologies for linguistics. LLKG is suitable for holding information from various collections by interconnecting the entities with external resources to expand the reachable knowledge. The resources involved in this study are the Linguistic Knowledge Graph, a time-sensitive graph database for linguistic knowledge, and Etymological WordNet, a lexical resource for describing word origins. Except for the significance of the data they hold, these resources were considered aligned with the objectives of this work due to their design choices, which hinder the contribution to the cloud of Linguistic Linked Open Data and prevent the discovery of new knowledge. Figuratively speaking, we intend to burst the bubble of their isolation. To this purpose, this work focuses on translating the Labelled Property structure into Resource Description Framework Schema and adopting the lexicon model for ontologies LEMON. This work also illustrates how to enrich the graph by manually linking its entities to resources on the Web, e.g., Universal WordNet, LiLa and Wikidata.

1 Introduction and Motivations

Words originate and evolve in several scenarios: different pronunciations diverge in different languages; changes occur within the same language as historical events unfold; distinct languages come into contact and borrow words from one another. Similarly, the meanings associated with words should not be regarded as static over time, as they are dynamic and constantly evolving to reflect cultural changes. This vitality of language causes the semantics of words to undergo considerable mutations (Traugott, 2005), ranging from an utter transformation of their core to a slight shift; in particular, they can experience *pejoration* or *amelioration* when meanings become respectively more negative or more positive, or they can *broaden* or *narrow* their scope. Such mutations are often tightly correlated with the culture of the period they occur in. Consider the Latin adjective *beatus*, annotated in McGillivray et al. (2022) with five senses: "happy", "fortunate", "rewarded", "rich" and "blessed". When considering the period in which these senses occur, it is noticeable that the sense "blessed" only emerged later with the advent of Christianity, overshadowing the other senses.

Recent years have seen a rising interest in computational *Lexical Semantic Change Detection* (LSCD) (Basile and McGillivray, 2018; Tsakalidis et al., 2021; Kutuzov et al., 2018; Tahmasebi et al., 2021; Castano et al., 2022); the availability of large corpora and the development of computational semantics have stimulated numerous initiatives for capturing semantic change in a data-driven fashion. In Armaselu et al. (2022), authors have encouraged the integration of distributional approaches for Natural Language Processing (NLP) with Linked Open Data[1] (LOD) technologies, emphasising how these external resources better support the heterogeneous nature of the data relevant to this phenomenon, which include linguistic knowledge but also information on historical events and entities, as well as bibliographical and geographical data. Particularly, such technologies could be leveraged in linguistics to investigate word histories, conduct etymological research, and analyse quantitative patterns in the distribution of word senses across time and according to the authors of the texts and other textual features. The knowledge attainable from external resources can also be valuable in explaining the motives behind a change in semantics, which is often tightly bound to social and cultural aspects. This type of information can not be acquired via observing and analysing a corpus, but requires the support and involvement of external knowledge. LOD are

[1] w3.org/DesignIssues/LinkedData

based on ontologies and Knowledge Graphs (KG) (Hogan et al., 2021), graphs of data intended to convey knowledge of the real world, allowing the modelling of complex domains. Knowledge Graphs offer several advantages: *(i)* they hold semantically structured, potentially very large, machine-readable data collections; *(ii)* they can be enhanced with schemas and ontologies to define and reason about the semantics of nodes and edges, making it possible to discover implicit knowledge; *(iii)* a KG and its schema are not language-specific; *(iv)* LOD principles encourage interlinking ontologies and data, therefore it is possible to discover even more knowledge by traversing these external links. Despite these advantages, integrating LOD to help diachronic analysis has not yet been thoroughly investigated.

This work intends to *(RQ1)* design a knowledge graph for diachronic analysis *(i)* set on well-grounded ontologies for linguistics, i.e., LEMON, and *(ii)* fit for holding information coming from diverse datasets; *(RQ2)* link the entities to external resources to enrich knowledge.

The paper is structured as follows: Section 2 introduces the resources at issues while Section 3 illustrates in detail the design and the linking process behind the LLKG. Finally, Section 4 summarises our contribution and proposes future works.

2 Resources

This section briefly illustrates the resources involved in the construction of the LLKG.

2.1 Etymological WordNet

Etymological WordNet[2] (De Melo, 2014) (EtymWN) is a lexical resource for describing word origins as relationships between two terms, even from different languages, in a machine-readable network. It is intended as a network of words that can capture etymological and word synchronic and diachronic information in a lexical network; however, no word sense-specific information is considered. Relations are reported in a triple format, $\langle \texttt{lang}_a\texttt{:t}_1\texttt{, rel:relation, lang}_b\texttt{:t}_2 \rangle$, where $\texttt{t}_1$ and $\texttt{t}_2$ are terms, $\texttt{lang}_a$ and $\texttt{lang}_b$ are (not necessarily) different languages.

The EtymWN graph has been mined from the 2013-09-07 version of English Wikitionary, obtaining a network of 3,000,000 terms. EtymWN models 500,000 etymological origin links, 500,000 etymological relatedness links, and 2,300,000 derivational and compositional links. Thanks to the graph representation, EtymWN makes navigating and uncovering connections between words, even unexpected ones, much more explicit; besides, network-like graphs are machine-readable and language-neutral. Nevertheless, EtymWN Achilles' heel is its own source: as a matter of fact, Wikitionary allows *anyone* to contribute, thus it is sensible to suppose that some contributions might be inaccurate or false. That is why both Wikitionary and Etymological WordNet are not to be considered indisputable sources, still they are very useful as exploratory tools.

2.2 Linguistic Knowledge Graph

Linguistic Knowledge Graph (Basile et al., 2022) (LKG) aims at capturing different aspects of lexical resources, such as relations between words and concepts, morphological and syntactical information. Moreover, it can cover diachronic aspects of language: the date of publication of a document and the birth and death of an author. The LKG models time-sensitive linguistic knowledge using a graph database. Its purpose is to lay the foundations for the study of word histories, for etymological research and, finally, for the analysis of the distribution of word senses not only over time but also according to the authors of the text and other textual features.

The LKG schema was designed by taking inspiration from the ontology-lexicon model LEMON (McCrae et al., 2012) and semantic networks such as WordNet and BabelNet. This graph model has been developed with the intent of modelling *(i)* relations between concepts and words, *(ii)* information about word occurrences, *(iii)* diachronic information of concepts and words. Conversely to LEMON built on RDF-S and OWL, the structure of LKG is based on the Labelled Property Graph (LPG) model, ensuring great flexibility and expressive power. In LPG both nodes and arcs are associated with unique identifiers, can be labelled and can store property values as `attribute-value` maps. Section 3 will illustrate the details of the process for converting the schema from LPG to RDF-S.

2.2.1 Data sources

LKG imports (McGillivray et al., 2023a,b) a portion of the LatinISE corpus (McGillivray and Kilgarriff, 2013; McGillivray et al., 2022), which gathers ten million word tokens that have been lem-

[2] `etym.org`

matised and PoS-tagged. In addition, 40 selected Latin lemmas were included, of which 17 have undergone a semantic shift and 23 have maintained their original meaning. For each lemma, 60 fragments were randomly extracted from the corpus, 30 dated BCE and 30 CE, and have been manually annotated by 10 scholars with a high-level knowledge of Latin, following the DuReL framework (Schlechtweg et al., 2018) which measures the semantic relatedness of a word usage with respect to its dictionary definitions. The motivations behind the choice of LatinISE are three: *(i)* data in LatinISE are not compliant with the Linked Data principles, therefore we took on the challenge of translating it from Labelled Property Graph to Resource Description Framework; *(ii)* besides manual annotated senses, LatinISE includes metadata for each fragment valuable in a diachronic analysis setting, i.e., the author, his or her date of birth and death, and occupation, the opus it occurs in and its publishing date; *(iii)* LatinISE is the dataset of choice for Latin in the SemEval-2020 Task 1 (Schlechtweg et al., 2020), in view of assessing the contribution of LLKG to the lexical semantic change detection task.

In a Lexical Semantic Change setting, it is crucial to have a language-specific model available, as well as access to extensive corpora covering several periods. Among historical languages, Latin is one of the most represented thanks to several factors: *(i)* accessible digital data covering two thousand years of history, e.g., LiLa (Passarotti et al., 2020, 2019b), Latin WordNet (Minozzi, 2017) and LatinISE, *(ii)* extensive computational language resources specially designed for Latin are available, e.g., Classical Language Toolkit (Johnson et al., 2021), UDPipe (Straka et al., 2016; Straka and Straková, 2017), *(iii)* ancient languages offer the opportunity to study long-term lexical semantic change and Latin itself is a prime example of a language that is not only ancient but has also continued to be actively used long after the end of antiquity, undergoing various diachroinc evolution observable in a wealth of textual data (Stroh, 2007; Leonhardt, 2013).

Refinements. After comparing the schema and the graphical representation followed by a qualitative analysis of the dataset, several discrepancies were revealed, especially in the naming of classes, relations and attributes. Moreover, a few relations and classes from the schema are missing from the dataset and vice versa. A comprehensive and compact list mapping schema, graph and dataset is available in the project documentation[3].

It was noticed that eleven lemma nodes had an incorrect part-of-speech tag: they were labelled as nouns whilst, according to the list provided in (McGillivray et al., 2022), they are verbs or adjectives. To avoid downstream inaccuracies, the PoS-tags of the following words were manually corrected: *acerbus*, *adsumo*, *beatus*, *credo*, *dubius*, *fidelis*, *itero*, *licet*, *necessarius*, *oportet*, *simplex*. Finally, it was observed that four lemmas in the dataset were spelled differently, once again w.r.t. the above mentioned list: *civitas*, *jus*, *virtus* and *voluntas* were replaced with *ciuitas*, *ius*, *uirtus* and *uoluntas*, respectively.

Integration. Due to the incompleteness of the author's mapping to Wikidata, 586 sentences out of 2,398 were missing. After a manual mapping, all the 586 missing sentences, together with 82 authors, respective Wikidata entities, and 114 works, have been included in the dataset; if an author Wikidata entity was not available, `None` was used. There are cases in which the author is uncertain or unknown, but these notations were not shared among annotators (e.g., `unknown`, `[Auctor incertus]`, `No Author`, `[Anonymous]`), therefore every different notation would result in a different entity in the graph. In order to unify these notations placeholder entities `Unknown author`, `Uncertain author`, `Anonymus`, and `Various authors` have been introduced. See Table 1.

Entity	**Count**
`Unknown author`	79
`Uncertain author`	81
`Anonymus`	10
`Various authors`	14

Table 1: Count of each placeholder entity, for a total of 184.

Additionally, during the mapping process, it became clear that many authors were occurring in the dataset with different notations, leading to multiple separated entities in the graph when there should be just one instead. Examples are "Plautus" and "Plautus Titus Maccius", "Ovidius", "Ouidius" and "Ovidius Naso Publius", which have been unified in "Plautus Titus Maccius" and "Ovidius Naso Publius", respectively.

[3] https://anonymous.4open.science/r/LLKG-2C87/

On the other hand, there are distinct works presenting the exact same title. Consider, for instance, Ovidius's "Metamorphoses", 8 AD, and Apuleius' "Metamorphoses", 2 AD. This conflict has been solved using one of the many alternative titles of Apuleius' work, "Asinus aureus".

The code for integrating the original dataset with missing authors and fragments is available on GitHub[4].

3 Linked Linguistic Knowledge Graph

The principal objectives of *Linked Linguistic Knowledge Graph* (LLKG) are to *reorganise* and *link* the contents of LKG (§2.2) and EtymWN (§2.1). Although existing semantic networks and ontologies heavily inspire the design of LKG, it defines its custom schema instead of reusing an already existing one violating the Linked Data principles[5]. Similarly, EtymWN does not provide a reusable definition of its relations. As a consequence, these design choices hinder the contribution of the resource to the Web of Linked Data, in particular to the cloud of Linguistic Linked Data (Chiarcos et al., 2011, 2012), and, conversely, prevent the discovery of new knowledge. To put this figuratively, we intend to burst the bubble of LKG and EtymWN.

As for the former objective (3.1), this work focuses on the translation of the LPG structure into RDF-S, in conformity with the state-of-the-art method described in (Hogan et al., 2021) and adopts the lexicon model for ontologies LEMON; on the other hand, entities have been manually linked to a variety of resources on the Web (3.2). From a technical perspective, the RDFLib 7.0.0 Python package[6] was used to generate a graph in Turtle syntax. The resource is freely available on Zenodo[7].

3.1 Schema

The schema has been informally divided into five sub-graphs, namely LINGUISTIC, EXAMPLE, AUTHOR, CORPUS, DATE. The figures of the graph, a detailed report of the mapping of each sub-graph from LPG to RDF-S and the employed vocabularies are available on GitHub[8]. The following sections highlight the necessary mappings from LKG to LLKG for converting the LPG format to RDF-S in the schema description.

3.1.1 Linguistic Sub-graph.

LEMON, short for *lexicon model for ontologies*[9] (McCrae et al., 2012), makes an effort to provide rich linguistic grounding for ontologies; it includes the representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface[10], i.e., the meaning of these lexical entries with respect to an ontology or vocabulary.

Traditional vocabularies like those offered by OWL and RDF-S do not support the definition of linguistic and lexical entities and relationships such as inflected forms, varied genders, usage notes, or the creation of a comprehensive lexical resource due to their general nature. Consequently, LEMON aims to bridge this gap by providing a vocabulary that enriches ontologies with linguistic details.

Due to the conformation with the LEMON model and the underlying translation from an LPG to RDF-S structure, the schema of the LINGUISTIC sub-graph has undoubtedly experienced the most substantial alterations with respect to its original structure. This structure adaptation is mainly grounded on ONTOLEX and VARTRANS modules: the former models the lexical entries and senses, while the latter introduces vocabulary for representing relations between them.

LEXICALENTRY. Starting from the core of this sub-graph, the lexical entry is represented in LKG by the `LexiconEntry` class, which can be specialised into `Lemma` or `WordForm`, suggesting that lemmas and entries are in some way related. On the other hand, LEMON separates these entities: `ontolex:LexicalEntry` has three sub-classes, `ontolex:Word`, which corresponds to `WordForm` in LKG, `ontolex:Affix` and `ontolex:MultiwordExpression`, whereas a lemma is an `ontolex:Form` entity linked to an `ontolex:LexicalEntry` through the `ontolex:canonicalForm` relation; the actual lemma string is a literal connected to `ontolex:Form` via `ontolex:writtenRep`.

`ontolex:Form` represents one generic grammatical realisation of a lexical entry, thus its instance alone is not enough to represent the concept

[4] github.com/pippokill/GraphLatinISE

[5] w3.org/DesignIssues/LinkedData

[6] rdflib.readthedocs.io/en/stable/index.html#

[7] zenodo.org/records/14623212

[8] github.com/Midorilly/STKG-LLKG/

[9] w3.org/2016/05/ontolex/

[10] w3.org/2016/05/ontolex/#syntax-and-semantics-synsem

of lemma and requires ontolex:canonicalForm to translate the :HAS_LEMMA from LKG. As for the lemma properties posTag and mwe, they have been converted into two distinct entities according to LEMON; in LLKG, the former is a lexinfo:PartOfSpeech entity linked to ontolex:Form via lexinfo:partOfSpeech, the latter is represented with the aforementioned ontolex:MultiwordExpression.

ontolex:LexicalEntry is connected to dct:LinguisticSystem via dct:language; even though this structure may appear identical to LKG, it is to be noted that since Lemma is a sub-class of LexiconEntry, it can be connected to Language via :HAS_LANGUAGE. This is not reflected in LLKG, where only an ontolex:LexicalEntry is associated with its respective language entity. Properties of Language iso639-1 and iso639-2 in LKG have been transposed into relations iso6391 and iso6302 as sub-properties of dct:identifier connected to dct:LinguisticSystem; additionally, relation iso6393 was included.

:{LEX_RELATION} between two LexiconEntry has found its counterpart in vartrans:lexicalRel between two ontolex:LexicalEntries. As suggested in the vartrans documentation[11], it would be preferable to introduce more specific sub-properties in place of vartrans:lexicalRel, which relates two lexical entries that stand in some unspecified lexical relation; however, LKG lacks the necessary information.

As for Etymological WordNet, an extension of ONTOLEX-LEMON vocabulary for modelling etymology, LEMONETY[12], has been designed by Khan. However, the information in EtymWN does not provide a deep understanding of the etymologies (e.g., etyma, cognate forms, respective lemmas), which is instead required by LEMONETY; furthermore, considering *(i)* the reification of every etymological link, etymology and etymon necessary according to the schema of LEMONETY, and that *(ii)* EtymWN graph counts 3,000,000 terms and approximately 6,000,000 etymological relations, it would result in a very considerable number of additional nodes. Therefore, for the time being the following relations were defined: llkg:etymology, llkg:etymologicalOriginOf, llkg:etymologicallyRelated, llkg:hasDerivedForm, llkg:isDerivedFrom, llkg:orthographyVariant. They all relate to two ontolex:LexicalEntry entities and are sub-properties of vartrans:lexicalRel, except for llkg:orthographyVariant. Notwithstanding, in light of the *do not reinvent the wheel* cornerstone, LEMONETY offers a finer modelling solution to be considered in future works.

LEXICALSENSE. The second cornerstone of the LINGUISTIC sub-graph is represented by lexical senses. In LKG, they are LexiconConcept, whereas LLKG employs ontolex:LexicalSense. Lexical senses are word senses gathered into synsets, expressed in LKG as Concept and in LLKG as ontolex:LexicalConcept. The :REFER_TO link between a sense and a synset in LKG has been replaced with ontolex:isLexicalizedSenseOf. An additional relation introduced by LEMON, ontolex:evokes between a lexical entry and a concept was included in the schema.

Whilst LKG defines the :HAS_DEFINITION relation between LexiconConcept and Text to embody the gloss of the sense, LLKG employs dct:description between ontolex:LexicalSense and rdfs:Literal. This choice was made to define a more straightforward and disjoint schema since in LKG, the Text entity is used for representing both sense glosses and textual excerpts, resembling a datatype more than an actual class.

Although this is not specified in the LKG schema and graph visualisation, a SAME_AS relation occurs in the dataset; following a qualitative analysis, we concluded that it is used for relating equivalent senses from different resources, e.g., Latin WordNet and Lewis-Short Dictionary. LLKG adopts owl:sameAs. From the LatinISE dataset structure, it was clear that lexical entries are linked to senses extracted from the Lewis-Short Dictionary only, pointing to their Latin WordNet corresponding sense. Attribute resource was converted into relation dct:source linking ontolex:LexicalSense and rdfs:Resource. In the case of senses from Latin WordNet, two additional properties, llkg:wn30ID and llkg:wn31ID, were included to hold the WordNet 3.0 and 3.1 sense identifiers; a supplementary relation rdfs:seeAlso connects each sense with the LatinWordNet 3.1 URI re-

[11] lexinfo.net/ontology/2.0/lexinfo#

[12] github.com/anasfkhan81/lemonEty

trieved from LiLa (see Section 3.2.2).

As concerns relations between senses, e.g., hyponymy and hypernymy, they were expressed in LKG via a generic :{SEM_RELATION}, without providing further details. With the WordNet Interface provided by NLTK[13], it was possible to reconstruct the specific relation between synsets using their WordNet identifiers. Therefore, the LLKG schema leverages the wn:hypernym and wn:hyponym relations from the WordNet schema.

At last, the "glue" relation between words and their senses :HAS_CONCEPT has its LEMON counterpart in ontolex:sense linking an ontolex:LexicalEntry entity to one or more ontolex:LexicalSense entities.

3.1.2 Example Sub-graph

The EXAMPLE sub-graph is the joining link between the LINGUISTIC and CORPUS sub-graphs and contains crucial information for the study of semantic shift of words. An example is a fragment of text in which a specific word occurs with a particular sense.

The LKG schema does not specify an example as an entity but expresses this concept with the relation :HAS_EXAMPLE between LexiconConcept and Text, with attributes begin, end and grade; similarly, LKG expresses the concept of a word occurring in a text with the relation :HAS_OCCURRENCE between LexiconEntry and Text. In LLKG the originally merged usage of entity Text for representing both a fragment from a text and the actual definition of a word sense has been split, obtaining two distinct classes: wn:Example from the GLOBAL WORDNET RDF SCHEMA and schema:Quotation from SCHEMA.ORG, linked via dct:isPartOf. As concerns the :HAS_EXAMPLE relation, the WordNet Schema defines wn:example as "an example usage of a sense or synset", connecting an ontolex:LexicalSense with a wn:Example and making the translation rather direct. Conversely, no definition of a relation between a lexical entry and an example is available, therefore :HAS_OCCURRENCE has become dct:isPartOf from an ontolex:LexicalEntry to a wn:Example. Attributes indicating the offset of a word string in an example, begin and end of :HAS_OCCURRENCE and :HAS_EXAMPLE, were collapsed into two relations powla:start and powla:end, linking wn:Example to an unsigned integer. Considering that the offset of word sense corresponds to the offset of the word form conveying it, duplicating these two properties as LKG does would result in an unnecessary redundancy.

Finally, :HAS_EXAMPLE is characterised by the grade attribute, which carries the annotation score of a sense when its word form occurs in a fragment, essential information for diachronic analysis. In LLKG, grade is a relation llkg:grade between wn:Example and a rdfs:Literal of type float.

Below is an example of the lexical entry dubious.

```
<http://lexvo.org/id/term/eng/dubious> a
     ontolex:Word ;
    rdfs:label "dubious"^^xsd:string ;
    llkg:etymologicallyRelated <http://lexvo.
        org/id/term/eng/doubt>,
       <http://lexvo.org/id/term/eng/
           dubitation> ;
    llkg:etymology <http://lexvo.org/id/term/
        lat/dubius> .

<http://lexvo.org/id/term/lat/dubius> a
     ontolex:Word ;
    rdfs:label "dubius"^^xsd:string ;
    llkg:hasDerivedForm <http://lexvo.org/id/
        term/lat/dubiam> ;
    llkg:llkgID 4161225 ;
    dct:language <http://lexvo.org/id/iso639-3/
        lat> .

<http://lexvo.org/id/term/lat/dubiam> a
     ontolex:Word ;
    rdfs:label "dubiam"^^xsd:string ;
    llkg:isDerivedFrom <http://lexvo.org/id/
        term/lat/dubius> ;
    dct:isPartOf llkg:example_4102 ;
    dct:language <http://lexvo.org/id/iso639-3/
        lat> ;
    ontolex:canonicalForm <http://lila-erc.eu/
        data/id/lemma/100177> ;
    ontolex:sense llkg:dubius-0, llkg:dubius-1,
         llkg:dubius-2 .

<http://lila-erc.eu/data/id/lemma/100177> a
     ontolex:Form ;
    rdfs:label "dubius"^^xsd:string ;
    llkg:llkgID 4039 ;
    lexinfo:partOfSpeech lexinfo:adjective ;
    ontolex:writtenRep "dubius"@la .
```

3.1.3 Date Sub-graph.

Plenty of schemata are already available for modelling persons and corpora entities; for the DATE, AUTHOR and CORPUS sub-graphs we opted for SCHEMA.ORG because of its considerably extended vocabulary, which made the translation quite direct.

LKG concretely distinguishes a TemporalSpecification in sub-classes TimePoint and TemporalInterval; the former contains the actual information about a

[13] nltk.org/

date, i.e., attributes year, month and day, the latter is related to two TimePoint nodes via startTime and endTime relations. LLKG leverages schema:Date datatype for representing both TimePoint and TemporalInterval; dates were converted according to the ISO-8601 standard date format as required by SCHEMA.ORG: YYYY-MM-DD for TimePoint and YYYY-MM-DD/YYYY-MM-DD for TemporalInterval. As a result, in LLKG schema:Date is intended as the datatype of rdfs:Literal, not a class.

3.1.4 Author Sub-graph.

LKG Person is now schema:Person, and the properties name and surname are relations schema:givenName and schema:familyName, respectively. :BORN and :DIED are schema:birthDate and schema:deathDate. Although it is not specified in the schema of LKG, nodes of class Occupation and relation HAS_OCCUPATION occur in the dataset. To this purpose, the schema:Occupation class and the schema:hasOccupation relation were employed. Additionally, the schema:Organization class was included in order to model the concept of corpora authors, which usually are research teams or organisations.

3.1.5 Corpus Sub-graph.

Finally, Text, Document and Corpus were transposed into schema:Quotation, schema:Book and schema:Collection respectively, and they are all sub-classes of schema:CreativeWork. As mentioned earlier, schema:Quotation is the result of the split of Text class: this means that for each Text node in LKG there are a schema:Quotation and a wn:Example nodes in LLKG, connected by dct:isPartOf. schema:Quotation is the class that actually holds the string of the fragment via schema:text. LKG Sentence sub-class of Text was discarded since we felt this was an unnecessary specialisation. Relation :BELONG_TO between Text and Document, Document and Corpus was translated into schema:isPartOf.

As for the :HAS_AUTHOR relation, SCHEMA.ORG provides schema:author linking a schema:CreativeWork and a schema:Person or schema:Organization; in LKG, all three classes are linked to their author. However, we found that specifying the author of both a fragment and the document it belongs to is excessive; consequently, in LLKG only schema:Book and schema:Collection utilise the schema:author relation since they most likely differ. :PUBLISHED_IN is now schema:datePublished. All three classes are also connected to dct:LinguisticSystem via dct:language, as in LKG.

3.2 Linking

After determining the resources of interest, we settled on proceeding with a manual linking via SPARQL endpoints. This section briefly describes the resources and illustrates the queries. Below an overview of which classes are linked to which resource: *Lexvo.org* for ontolex:LexicalEntry and dct:LinguisticSystem; *Universal WordNet* for ontolex:LexicalSense; *LiLa* for ontolex:Form and ontolex:LexicalSense rdfs:seeAlso; *Wikidata* for ontolex:Form rdfs:seeAlso, schema:Person, schema:Occupation and schema:Book.

3.2.1 Lexvo.org and Universal WordNet

Lexvo.org (de Melo, 2015) is a service that publishes information about numerous aspects of human language online in both human-readable and machine-readable form, contributing to the Web of Linked Data and the Semantic Web. It defines URIs for terms, languages, scripts, and characters, which are not only highly interconnected but also linked to a variety of resources on the Web. Lexvo.org also includes Universal WordNet (UWN)[14] (de Melo and Weikum, 2009). UWN is a large knowledge graph that aims at describing words, entities, and concepts in over 200 different languages in a large network structure. For each entry, UWN provides a corresponding list of meanings and shows how such meanings are semantically related. UWN includes the Princeton's WordNet lexical database (Fellbaum, 1998).

For instancing ontolex:LexicalEntry nodes, the URI of lexical entries is obtained by concatenating lexvo.org/id/term/ with the ISO 6391-3 code of the language and the entry itself, for example: ⟨lexvo.org/id/term/lat/dubius a ontolex:Word⟩.

As for ontolex:LexicalSense entries of Latin WordNet, to lexvo.org/uwn/entity/s/ was appended a string with the PoS-tag the WordNet 3.0

[14] mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/uwn

identifier: ⟨lexvo.org/uwn/entity/s/a960629 a ontolex:LexicalSense⟩.

Regarding the dct:LinguisticSystem nodes, Lexvo.org provides a dump from which the URIs of entities of type lvont:Language, the preferred label and ISO 639-1, 639-2 and 639-3 codes (if defined) were fetched.

3.2.2 LiLa

The LiLa project[15] (Passarotti et al., 2020, 2019b) has built a Linked Data-based Knowledge Base of Linguistic Resources and Natural Language Processing tools for Latin. The core of LiLa is an extensive collection of Latin lemmas that serves as its foundation. Interoperability among the resources is facilitated by connecting all lexical resource entries and corpus tokens to a common lemma. LiLa also offers a SPARQL endpoint for access[16].

To fetch the URI of a Lemma node, its written representation ?written and its PoS-tag ?pos are bound to its value and posTag attributes, respectively. Since the written representation alone might point to more than one lemma, the PoS-tag is leveraged to disambiguate. Nevertheless, there are some cases in which the PoS-tag is not enough, e.g., salus: the query returns two lemmas, lila-erc.eu/data/id/lemma/123273, 2nd masculine declension of *salus, -i*, “high tide; sea; wave”, and lila-erc.eu/data/id/lemma/123276, 3rd feminine declension of *salus, -utis*, “health; salutation; Salvation”. Only if manually analysed it is possible to notice that the lexical entries related to the lemma *salus* in the dataset are declined forms of *salus, -utis*. Therefore, additional information such as the inflection type and the gender would make the query results more accurate. Currently all the returned URIs are retained.

```
SELECT ?lemma
WHERE {
    ?lemma ontolex:writtenRep ?written ;
    lila:hasPOS ?pos .
}
```

Regarding the sense URIs, if the attribute resource of LexiconConcept is “Latin WordNet”, it is possible to retrieve its WordNet 3.0 and 3.1 alpha-numeric identifier by giving in input the alias attribute, e.g., “sour.a.02”, to the NLTK WordNet interface, obtaining “02368787-a” and “02377355-a” respectively. In order to have an actual URI to bind ?lemmaURI to, this query is performed later on in the graph creation, after all the nodes and relations ontolex:canonicalForm and ontolex:sense have already been created, allowing us to navigate the graph.. ?resource is bounded to lila-erc.eu/data/lexicalResources/LatinWordNet/Lexicon. Finally, only the ?senseURI containing the aforementioned WordNet 3.1 identifier is retained.

```
SELECT ?senseURI
WHERE {
    ?resource lime:entry ?lexentry.
    ?lexentry ontolex:canonicalForm ?
        lemmaURI ;
    ontolex:sense ?senseURI .
    FILTER(regex(?senseURI,"{}")). }
}
```

In the case of attribute resource having value “Lewis-Short Dictionary,” retrieving a URI was not possible: the authors of LatinISE performed a manual selection of the senses, in some cases simplifying the gloss and not providing identifiers.

3.2.3 Wikidata

Wikidata is a collaborative, multilingual knowledge graph. Serving as a shared repository of open data, it is freely accessible to everyone and can be queried via its SPARQL endpoint. Wikidata is a document-oriented database whose pivotal elements are the items encapsulating topics, concepts, or objects. Although mapping of authors and occupations to Wikidata items is provided with the LatinISE dataset, not all the authors occurring in the dataset were mapped. Therefore, an attempt was made to fill possible gaps. The listings below illustrates the queries.

The first query retrieves the Wikidata Lexeme URI ?lexeme of a lemma, given its LiLa URI mentioned in 3.2.2, wdt:P11033 ?lila.

```
SELECT ?lexeme
WHERE {
?lexeme a ontolex:LexicalEntry ;
    wdt:P11033 ?lila .
FILTER(regex(?lila,"{}"))
}
```

As concerns the query for the author, first and foremost, ?authorURI must be an instance (wdt:P31) of human (wd:Q5); then rdfs:label or skos:altLabel should match ?label because it was noticed that in many cases the name of the author provided in the dataset, ?label, does not match the label of the Wikidata item.

[15] lila-erc.eu

[16] lila-erc.eu/sparql/lila_knowledge_base/sparql

```
SELECT ?authorURI
WHERE {
   ?authorURI wdt:P31 wd:Q5 .
   { ?authorURI skos:altLabel ?label .}
   UNION
   { ?authorURI rdfs:label ?label . }
FILTER (regex(str(?label), "{}", "i"))
} LIMIT 1
```

For instance, ⟨`wd:Q2039 a schema:Person`⟩ refers to "Titus Livius".

On the other hand, for obtaining a `?documentURI`, instead of looking at all the instances of written work (`wd:Q47461344`) and then filtering by label, an operation that might take too much time, the SPARQL pattern takes advantage of the information about the author to narrow the field. In this query, the author (`wdt:P50`) is bound to the `?authorURI` value retrieved from the graph. Since in many cases the `title` values of `Document` nodes correspond to the original Latin title, they rarely match with the preferred label of the Wikidata corresponding item. Therefore, likewise author query, a check for both `rdfs:label` and `skos:altLabel` to match `?label` is performed. However, several items correspond to this pattern; we are interested only in the original work. To avoid this situation, the retrieved URIs are limited to the ones having the same language (`wdt:P407`) as the author's language (`wdt:P6886`).

```
SELECT ?documentURI ?languageISO
WHERE {
   BIND (wd:{} AS ?authorURI)
   ?documentURI wdt:P50 ?authorURI .
   ?authorURI wdt:P6886 ?language.
   ?language wdt:P220 ?languageISO .
   { ?documentURI skos:altLabel ?label ;
      wdt:P407 ?language }
   UNION
   { ?documentURI rdfs:label ?label ;
      wdt:P407 ?language }
FILTER (regex(str(?label), "{}", "i"))
} LIMIT 1
```

For instance, ⟨`wd:Q1155892 a schema:Book`⟩ refers to "Ab Urbe condita" written by "Titus Livius", whilst the actual label of the Wikidata item is "History of Rome".

4 Conclusions and Future Work

We have introduced the *Linked Linguistic Knowledge Graph* (LLKG), a linguistic resource for supporting the diachronic analysis of language, reorganising and linking the contents of Linguistic Knowledge Graph and Etymological WordNet.

Etymological WordNet is a lexical resource that undertakes the effort of describing word origins in terms of relationships between two terms, even from different languages, in a machine-readable network. *Linguistic Knowledge Graph* captures different aspects of lexical resources, such as relations between words and concepts, morphological and syntactical information. Moreover, it covers diachronic aspects of language: the date of publication of a document, the birth and death of an author.

To this purpose, we have focused on translating the Labelled Property Graph structure into Resource Description Framework Schema and have adopted the lexicon model for ontologies LEMON. Etymological WordNet and a portion of LatinISE were loaded into our resource, and missing fragments have been integrated. Upon completion, the LLKG includes 194 authors, 406 literary works, 108 occupations, 527 senses, 7,908 languages and 2,879,193 lexical entries. Regardless of the data LLKG contains at the moment, it is important to highlight that the resource is language independent.

On the other hand, entities have been linked to various resources on the Web, e.g., Wikidata, resulting in more than 2,897,000 unique external links. Missing or incorrect Wikidata identifiers of authors were manually adjusted.

Future works include analysing the lexical and semantic relations in the LLKG to define more specific ones and integrating information about the lemmas, lexical entries, and fragments for more precise disambiguation when querying external sources. The necessity for a deeper analysis of etymological relations arose, together with the necessity of a more fine-grained vocabulary to represent them, i.e., LEMONETY (Khan, 2018). In addition to this, the integration of other annotated corpora, e.g., the LASLA Corpus and the Index Thomisticus Treebank already imported in LiLa (Fantoli et al., 2022; Passarotti et al., 2019a), can only benefit the LLKG resource.

Finally, studies on semantic change can leverage knowledge into LLKG to improve performance and provide explainability capabilities.

References

Florentina Armaselu, Elena-Simona Apostol, Anas Fahad Khan, Chaya Liebeskind, Barbara McGillivray, Ciprian-Octavian Truică, Andrius Utka, Giedrė Valūnaitė Oleškevičienė, 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 co-located with the 21st International Conference of the Italian Association for Artificial Intelligence (AIxIA 2022)*, page 69. CEUR Workshop Proceedings.

Pierpaolo Basile and Barbara McGillivray. 2018. Exploiting the web for semantic change detection. In *Discovery Science: 21st International Conference, DS 2018, Limassol, Cyprus, October 29–31, 2018, Proceedings 21*, pages 194–208. Springer.

Silvana Castano, Alfio Ferrara, Stefano Montanelli, Francesco Periti, et al. 2022. Semantic shift detection in vatican publications: a case study from leo xiii to francis. In *CEUR WORKSHOP PROCEEDINGS*, volume 3194, pages 231–243. CEUR-WS.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2011. Towards a linguistic linked open data cloud: The open linguistics working group. *Traitement automatique des langues*, 52(3):245–275.

Christian Chiarcos, Sebastian Hellmann, Sebastian Nordhoff, Steven Moran, Richard Littauer, Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer. 2012. The open linguistics working group. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3603–3610, Istanbul, Turkey. European Language Resources Association (ELRA).

Gerard De Melo. 2014. Etymological wordnet: Tracing the history of words. In *LREC 2014*, pages 1148–1154.

Gerard de Melo. 2015. Lexvo.org: Language-related information for the Linguistic Linked Data cloud. *Semantic Web*, 6(4):393–400.

Gerard de Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 513–522, New York, NY, USA. Association for Computing Machinery.

Margherita Fantoli, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. Linking the LASLA corpus in the LiLa knowledge base of interoperable linguistic resources for Latin. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 26–34, Marseille, France. European Language Resources Association.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutiérrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. *Knowledge Graphs*. Number 22 in Synthesis Lectures on Data, Semantics, and Knowledge. Springer.

Kyle P. Johnson, Patrick J. Burns, John Stewart, Todd Cook, Clément Besnier, and William J. B. Mattingly. 2021. The Classical Language Toolkit: An NLP framework for pre-modern languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 20–29, Online. Association for Computational Linguistics.

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

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*.

Jürgen Leonhardt. 2013. *Latin: Story of a world language*. Harvard University Press.

John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, et al. 2012. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46:701–719.

Barbara McGillivray, Pierluigi Cassotti, Pierpaolo Basile, Davide Di Pierro, and Stefano Ferilli. 2023a. Using graph databases for historical language data: Challenges and opportunities. In *Proceedings of the 19th Conference on Information and Research Science Connecting to Digital and Library Science (IRCDL 2023)*.

Barbara McGillivray, Pierluigi Cassotti, Davide Di Pierro, Paola Marongiu, Anas Fahad Khan, Stefano Ferilli, and Pierpaolo Basile. 2023b. Graph databases for diachronic language data modelling. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 86–96, Vienna, Austria. NOVA CLUNL, Portugal.

Barbara McGillivray and Adam Kilgarriff. 2013. Tools for historical corpus research, and a corpus of latin. *New methods in historical corpus linguistics*, 1(3):247–257.

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.

Stefano Minozzi. 2017. Latin wordnet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'information retrieval. *Strumenti digitali e collaborativi per le Scienze dell'Antichita*, (14):123–134.

Marco Passarotti, Eleonora Litta, Flavio Massimiliano Cecchini, Matteo Pellegrini, Giovanni Moretti, Paolo Ruffolo, and Giulia Pedonese. The lila knowledge base of interoperable linguistic resources for latin. architecture and current state.

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(1):177–212.

Marco Passarotti et al. 2019a. The project of the index thomisticus treebank. *Digital classical philology. Ancient Greek and Latin in the digital revolution*, 10:299–319.

Marco Carlo Passarotti, Flavio Massimiliano Cecchini, Greta Franzini, Eleonora Litta, Francesco Mambrini, and Paolo Ruffolo. 2019b. The lila knowledge base of linguistic resources and nlp tools for latin. In *LDK (Posters)*, pages 6–11.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. *Preprint*, arXiv:2007.11464.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (durel): A framework for the annotation of lexical semantic change. *arXiv preprint arXiv:1804.06517*.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, pages 88–99.

Wilfried Stroh. 2007. Latein ist tot, es lebe latein!: kleine geschichte einer grossen sprache. *(No Title)*.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, 6(1).

Elizabeth Closs Traugott. 2005. Semantic change: Bleaching, strengthening, narrowing, extension. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*, pages 124–31. Elsevier.

Adam Tsakalidis, Pierpaolo Basile, Marya Bazzi, Mihai Cucuringu, and Barbara McGillivray. 2021. Dukweb, diachronic word representations from the uk web archive corpus. *Scientific Data*, 8(1):269.