A Dataset of Latin Etymologies Extracted from Wiktionary

Javier de Torres
Independent Researcher
Madrid, Spain

Marco Passarotti
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Milan, Italy

Francesco Mambrini
CIRCSE Research Centre
UCSC
Milan, Italy

Matteo Pellegrini
Surrey Morphology Group
University of Surrey
Guildford, United Kingdom

Giovanni Moretti
CIRCSE Research Centre
UCSC
Milan, Italy

Abstract

We present a curated resource of Latin etymologies automatically extracted from Wiktionary, enriched with links to the LiLa Knowledge Base of Latin and modelled as RDF triples using the LemonEty ontology. We also present the Python pipeline the data was generated with, as it can be reused to extract Wiktionary's etymologies for other languages. The etymology chains cover Latin words and their attested or reconstructed ancestors in languages such as Proto-Indo-European, Proto-Italic, Ancient Greek, Hebrew, Egyptian, and others. To address the structural noise and editorial heterogeneity of Wiktionary etymology data, we have introduced strong rule-based filters throughout the pipeline, especially in the curation stage. After validation, the resulting dataset contains etymological chains for 9,684 lemmas, which can be used to support research in Historical Linguistics, Computational Etymology and language learning, among other applications.

1 Introduction

Etymological information is valuable for a wide array of applications related to lexicography, including Historical Linguistics and the construction of knowledge graphs for lexical resources, such as large multilingual semantic networks like BabelNet, which integrates lexical and encyclopedic knowledge across languages (Navigli and Ponzetto, 2012), and linguistic knowledge graphs that model relations between lexical units and concepts for digital humanities research (Basile et al., 2022). However, high-quality structured etymological datasets remain relatively scarce. Wiktionary contains a large collection of high-quality etymological descriptions contributed by its community of editors. However, these descriptions are primarily expressed in unstructured natural language. Our initial goal was to create software that took them as input and returned them in a structured format as output.

To embody our objective in a practical application, we focus our development on enriching the collection of lemmas that constitutes the core component of Linking Latin (LiLa)[1], a knowledge base of Latin linguistic resources made interoperable through their publication as Linked Open Data. To do this, we conduct the automatic extraction, normalization and structuring of Latin etymologies from Wiktionary into RDF format, and their subsequent alignment with LiLa lemmas. The main contributions of this work are (i) the resulting dataset of Latin etymological chains extracted from Wiktionary and (ii) the rule-based pipeline to generate that data. The dataset contains curated chains for 9,684 lemmas and, with respect to LiLa, provides a complementary layer of etymological information derived from a large-scale, collaboratively curated source.

In addition, the pipeline was designed following a clear separation-of-concerns principle, resulting in a highly modular workflow. Each stage of the process—extraction, enrichment, curation, and RDF serialization—can therefore be modified independently. This design makes the pipeline easily adaptable to other use cases. For instance, extracting etymologies for a different language and linking them to another knowledge graph would only require modifying the enrichment step preceding RDF serialization.

2 Related Work

Wiktionary has increasingly been used as a large-scale lexical resource for linguistic and NLP applications. Several efforts have focused on transforming its collaboratively edited content into structured data that can be processed automatically. In particular, the Wiktextract system (Ylonen, 2022) provides a comprehensive extraction pipeline capable of expanding Wiktionary templates and Lua mod-

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

ules, producing machine-readable representations of lexical entries and their associated linguistic information.

The Linked Open Data (LOD) paradigm promotes the publication of structured data on the Web using standard technologies such as RDF and HTTP identifiers, enabling datasets from different sources to be interlinked and reused across applications (Berners-Lee, 2006). Within the field of linguistics, this paradigm has encouraged the development of interoperable lexical resources published as Linked Open Data, allowing lexical information to be integrated into broader knowledge graphs and shared across linguistic infrastructures.

Within this context, the representation of lexical data as linked data has been extensively explored within the OntoLex-lemon framework (McCrae et al., 2017), a W3C community ontology designed to represent lexical information in RDF and to link lexical entries to their meanings and related linguistic data. In order to model etymological relations specifically, (Khan, 2018) proposed the lemonEty ontology, an extension of OntoLex-lemon designed to represent historical relationships between lexical items and their ancestors.

A number of projects have explored the publication of lexical data extracted from Wiktionary as Linked Open Data. One of the most prominent examples is DBnary (Sérasset, 2014), a large multilingual lexical resource automatically extracted from several language editions of Wiktionary and published in RDF using the lemon model. DBnary provides structured lexical information such as senses, translations, and morphological data, and has also served as a basis for experiments involving the extraction and visualization of etymological relations from Wiktionary. Unlike DBnary, which primarily focuses on the extraction of lexical information such as senses, translations and morphology from Wiktionary, the dataset presented in this paper specifically targets the reconstruction of etymological chains and their representation using the lemonEty ontology, with explicit links to the LiLa LemmaBank.

Related efforts have investigated the construction of explicit etymology graphs from Wiktionary data. For instance, the Etytree project (Pantaleo et al., 2017) extracts etymological relationships between lexical items from Wiktionary and represents them as navigable graph structures, enabling interactive exploration of word histories and cross-linguistic derivational relations. However, the primary goal of Etytree is the visualization and exploration of etymological networks rather than the publication of structured etymological data as Linked Open Data. In contrast, the approach presented in this paper focuses on the construction of a curated dataset of etymological chains represented in RDF using the lemonEty ontology and aligned with the LiLa Knowledge Base, with the aim of enabling reuse and interoperability within the Linked Open Data ecosystem for linguistic resources.

Within the context of Latin linguistic resources, the LiLa Knowledge Base of Linguistic Resources for Latin integrates multiple datasets through Linked Open Data principles. In particular, Mambrini and Passarotti (Mambrini and Passarotti, 2020) demonstrate how etymological information can be modeled in LiLa using the OntoLex-lemon ontology together with the lemonEty extension. The dataset presented in this paper follows the same modelling approach, ensuring interoperability with the existing LiLa infrastructure.

3 Source Data and Corpus Construction

3.1 Data extraction

In Wiktionary, each entry for a word has one or several etymology sections, depending on how many etymologies have been proposed for it. Expressing etymologies in natural language poses a series of challenges when it comes to extracting the desired information. However, no parser development work was necessary on our side, as Tatu Ylonen already made Wiktextract, "the first known extractor capable of expanding Wiktionary templates and Lua modules" (Ylonen, 2022, p. 1317). In this context, Wiktionary templates are reusable pieces of wiki markup used to encode structured linguistic information within entries, such as derivation relations, inherited forms, or borrowings between languages. For instance, the template {{inh|en|ang|niht}} encodes that the English word night is inherited from Old English niht.

Every few days, Wiktextract is run on the English Wiktionary and the data dump is posted on `https://kaikki.org`. We downloaded the compressed .gz version that was on the site on March 16th 2026 and used its templates as our starting point. Wiktionary data has considerable size, so the best course of action is to always extract the minimal subset we intend to operate on. Wiktionary encodes etymological relations through a family

of templates that describe inheritance, derivation, borrowing, and roots, so, in this case, from the initial Wiktextract dump, we extracted the etymology templates for Latin and stored them in a JSON file, in which a key is a lemma and the value is the list of etymology templates for that lemma. See, for example, the templates for frater:

```
[
{"name": "inh", "args": {"1":
  "la", "2": "itc-pro", "3":
        "*frātēr"}},
{"name": "inh", "args": {"1":
  "la", "2": "ine-pro", "3":
       "*bʰréh₂tēr"}}
]
```

It is worth noting that a word can have several such lists, which makes the phrase "etymology templates" ambiguous. From this point onwards, we will use "etymology templates list" to denote a single templates list, like the one just shown, and "etymologies" to denote a list of these lists.

3.2 Data transformation

In this step, we removed all extraneous information from the etymology templates and reduced them to the minimal representation required to encode an etymological relation: a chain in which each element consists of a word–language pair.

To do so, given a word's etymologies, we iterate over each etymology templates list. For each list, we then iterate over each individual template and then classify it based on its `name` field:

- **Relevant**: contains the primary ancestry templates observed in the data, including inheritance (`inh`), derivation (`der`), borrowing (`bor`), and root relations (`root`), together with their extended variants (e.g. `inh+`, `der+`, `bor+`). These templates directly encode a relation between a target form and a source form in another language. Since inheritance, derivation, and borrowing all represent a step in a lexical ancestry chain, we treat them uniformly during transformation.

- **Inheritance-like**: groups together templates that behave structurally as ancestry transitions, including additional borrowing-related templates such as `ubor` (unadapted borrowing) and `slbor` (semi-learned borrowing). Although their linguistic interpretation differs slightly, they all encode the same structural pattern: a source form in a source language that precedes the target lemma.

- **inheritance-like (lite)**: Wiktionary occasionally uses simplified or lightweight template variants such as `inh-lite`, `der-lite`, and `bor-lite`. These templates follow the same argument structure as their standard counterparts but contain less metadata.

If a template fits into any of the three classes, we then normalize it into a minimal ancestry node of the form `(form, language)`, in which `form` is the template's word's lemma and `language` is its language code. Applying this procedure to all templates of an entry yields a linear etymology chain in template order. Given *frater*'s templates, see its etymology chain:

```
[(frātēr, itc-pro), (bʰréh₂tēr,
          ine-pro)]
```

After processing all words' etymologies, etymology chains were obtained for 12,421 Latin words and stored as a JSON, in which every key is a lemma and the value is its etymology chains.

In Wiktionary, the same etymology templates may be repeated for a single lemma. This can occur for two main reasons. First, entries are organized into multiple sections (e.g., by part of speech), and Wiktextract may treat these sections as separate entries while preserving the same etymology. For example, *Februārius* appears both as an adjective ("of February") and as a proper noun ("February"), with both entries containing identical etymological templates. Second, a surface form may appear both as a lemma and as an inflected form of another word. For instance, *animalis* appears both as an adjective lemma and as the genitive singular of *animal*, again yielding duplicate etymology template sequences. In addition, due to the normalization process, distinct template sequences may collapse into the same ancestry chain when non-etymological templates (e.g., suffixation or gloss information) are discarded during transformation.

To avoid redundancy, duplicate chains are removed per lemma prior to curation. This step ensures that each lemma is associated with a set of unique etymological chains while preserving all distinct ancestry information. In total, 555 lemmas were found to contain duplicate etymology chains and were subsequently deduplicated.

The chains reveal that Latin lexical items directly descend from a wide range of languages and language stages. These include proto-language reconstructions such as Proto-Italic (`itc-pro`) and Proto-Indo-European (`ine-pro`); historical Indo-European languages such as Ancient Greek (`grc`), Oscan (`osc`), and Etruscan (`ett`); and languages involved in lexical borrowing, including Phoenician (`phn`), Hebrew (`hbo`), Egyptian (`egy`), and Akkadian (`akk`). The dataset also contains language variants and dialectal designations used in Wiktionary, for instance Doric Greek (`grc-dor`), Ionic Greek (`grc-ion`), and Koine Greek (`grc-koi`), as well as historical stages such as Vulgar Latin (`VL.`), Late Latin (`LL.`), and Medieval Latin (`ML.`). Beyond these immediate ancestors, the etymology chains also reveal more distant indirect ancestry from additional languages and proto-languages that do not appear as direct sources. These include further Indo-European branches and reconstructed stages such as Proto-Germanic (`gem-pro`), Proto-Slavic (`sla-pro`), Proto-Indo-Iranian (`iir-pro`), and Proto-Semitic (`sem-pro`), as well as historical languages such as Hittite (`hit`), Old Irish (`sga`), Coptic (`cop`), and Old Church Slavonic (`cu`). Together, these chains illustrate the wide diachronic and geographic span captured by the dataset, reflecting both deep Indo-European ancestry and later borrowing and transmission across multiple languages.

3.3 Data curation

Due to the collaborative nature of Wiktionary, and the diversity and heterogeneity among contributors, there is always bound to be some noise in its data. For this reason, a rule-based curation stage is used to filter potentially unreliable chains. In designing these filters, we prioritize precision over recall: the goal is to maximize the proportion of correct etymological relations (true positives) in the final dataset, even if this means discarding some valid chains. As a consequence, the curation process may introduce a higher number of false negatives, but it significantly reduces the presence of incorrect or noisy etymological links. This trade-off is considered preferable for a corpus intended to support linguistic research and knowledge graph construction, where data reliability is particularly important. Furthermore, false negatives can always be made into true positives (e.g. by an annotator), whereas the process to undo false positives is more cumbersome, especially if the data is published and associated with URIs.

The etymologies associated with each lemma are divided into the following classes after applying the curation filters:

- **Valid**: The chain passes all filters and is considered a valid etymological chain.
- **Empty forms**: The word form is an empty string or a hyphen.
- **Repeated language**: The same language appears twice within a single ancestry chain.
- **Annotated forms**: The word form contains parentheses, indicating the presence of gloss commentary.
- **Morpheme entries**: The word begins with a hyphen, indicating an affix rather than a standalone lexical form.
- **Enumeration forms**: The form contains a comma, suggesting that multiple lemmas were enumerated in the same string.
- **Markup contamination**: The word contains markup characters such as `<`, `>`, `[`, `]`, `{`, or `}`.

Table 1 reports the outcome of the curation procedure. Lemmas with valid etymologies constitute approximately 78% of the original dataset, while the remaining 22% are filtered out by the curation rules. Among the filtered classes, the category "repeated language" (7%) is the one most likely to contain false negatives, since lexical items may undergo internal derivation, semantic shift, or other developments while remaining within the same language. By contrast, etymologies containing empty forms (13%) are certain to be true negatives, since an etymology cannot be valid while missing etymons.

3.4 RDF serialization

Given a list of curated etymology chains, the final stage of the pipeline consists of serializing them into RDF triples. In order to represent the data in a semantically meaningful and interoperable way, an ontological vocabulary must be used to model the relations between lexical entries and their ancestors. For this purpose, we adopt the OntoLex-lemon Etymology extension (*lemonEty*), a vocabulary specifically designed for representing etymological relations in RDF (Khan, 2018). The lemonEty model extends the OntoLex-lemon ontology with classes

Category	Count
Valid	9,684
Empty forms	1,639
Repeated language	861
Annotated forms	192
Morpheme-only forms	34
Enumeration patterns	9
Markup artifacts	2

Table 1: Distribution of valid and filtered etymologies by lemma after rule-based curation.

and properties that capture typical elements of etymological description, such as etymons and etymological links. Using this model allows the extracted chains to be published as linked data and facilitates interoperability with other linguistic resources.

Furthermore, lemonEty is already used in the LiLa Knowledge Base. Adopting the same model therefore ensures compatibility between our dataset and the existing LiLa infrastructure. According to said model, an etymology has the following elements:

- **Lexical entry**: the lexical item whose etymology is being described (e.g., *lupus*). This is represented as an `ontolex:LexicalEntry` and linked to its etymology.
- **Etymology node**: an instance of `lemonEty:Etymology` that acts as the central object representing the etymological description of the lexical entry.
- **Etymon**: the historical lexical forms from which the word derives (e.g., reconstructed forms such as **luk^{w}os*). These are represented as instances of `lemonEty:Etymon`, which is in turn a subclass of `ontolex:Form`.
- **Etymological links**: instances of `lemonEty:EtyLink` that connect successive stages of the etymological chain.
- **Source and target relations**: each etymological link specifies an `etySource` (the ancestor form) and an `etyTarget` (the derived form).
- **Canonical forms**: lexical entries are linked to their canonical form of citation through the property `ontolex:canonicalForm`. In the OntoLex-lemon model, a lexical entry is associated with an `ontolex:Form` representing its canonical form (i.e., the lemma), while the written representation of that form is expressed through the property `ontolex:writtenRep`. In LiLa, lemmas are modeled as instances of the class `lila:Lemma`, which is defined as a subclass of `ontolex:Form`. Linking lexical entries to LiLa lemmas through `ontolex:canonicalForm` therefore constitutes the standard mechanism used to connect lexical resources to the LiLa knowledge base.

In our pipeline, this model is instantiated automatically from the curated etymology chains. Given such a chain, the RDF serialization procedure represents the word whose etymology is described as an `ontolex:LexicalEntry`, linked to a central `lemonEty:Etymology` node. All preceding historical stages are modeled as `lemonEty:Etymon` instances. The diachronic progression is encoded through a sequence of `lemonEty:EtyLink` nodes connecting each stage to the next by means of `lemonEty:etySource` and `lemonEty:etyTarget`. These links are attached to the etymology node with `lemonEty:hasEtyLink`, and the first one is marked with `lemonEty:startingLink`.

Before serializing each etymology chain, we perform a preliminary enrichment step in which each node is associated with external identifiers. First, a Wiktionary URL is assigned to every form by attempting to resolve the corresponding page in the English Wiktionary. The system generates candidate URLs based on the form and its language code, accounting for special cases such as reconstructed proto-language forms, and verifies their existence through HTTP requests. The first candidate returning a valid response is retained as the Wiktionary reference for that node. During RDF serialization, this URL is then attached to the corresponding lexical entry or etymon using the property `schema:url`.

For Latin forms, an additional linking step connects the lemma to the LiLa Knowledge Base. This is done by querying the LiLa Text Linker service endpoint[2], which returns a linking key identifying the corresponding LiLa lemma when a match is found (Passarotti et al., 2024). In cases where multiple candidate lemmas are returned (e.g., due to homography), the pipeline selects the first candidate. Candidates are returned sorted in ascending

[2] https://lila-erc.eu/LiLaTextLinker/processText

order by lemma ID (e.g., `lilaLemma:103739`), so this corresponds to selecting the candidate with the lowest lemma ID. The linking key is then converted into a persistent LiLa URI and used during RDF serialization to link the lexical entry to the corresponding LiLa resource through `ontolex:canonicalForm`.

Given the etymology chain of *frater* shown in Section 2.2, Listing 1 shows a simplified excerpt of its RDF representation. Prefixes such as `exlex`, `exety`, `exetym`, and `exlink` are used instead of full URIs for readability and stand for the corresponding namespaces in the dataset.

Listing 1: Simplified RDF triples for the etymology of *frater*.

```
exlex:frater-la a ontolex:LexicalEntry ;
  lemonEty:etymology exety:frater-la ;
  schema:inLanguage "la"^^xsd:language ;
  schema:url <https://en.wiktionary.org/wiki/
    frater> ;
  rdfs:label "frater" ;
  ontolex:canonicalForm <https://lila-erc.eu/
    data/id/lemma/103739> .

exety:frater-la a lemonEty:Etymology ;
  lemonEty:etymon exetym:frater-la-0 , exetym:
    frater-la-1 ;
  lemonEty:hasEtyLink exlink:frater-la-1 ,
    exlink:frater-la-2 ;
  lemonEty:startingLink exlink:frater-la-1 ;
  rdfs:label "Etymology of: frater" .

exetym:frater-la-1 a lemonEty:Etymon ;
  schema:inLanguage "x-ine-pro"^^xsd:language ;
  rdfs:label "*brhtr" .

exetym:frater-la-0 a lemonEty:Etymon ;
  schema:inLanguage "x-itc-pro"^^xsd:language ;
  rdfs:label "*frtr" .

exlink:frater-la-1 a lemonEty:EtyLink ;
  lemonEty:etySource exetym:frater-la-1 ;
  lemonEty:etyTarget exetym:frater-la-0 ;
  rdfs:label "Etymology Link" .

exlink:frater-la-2 a lemonEty:EtyLink ;
  lemonEty:etySource exetym:frater-la-0 ;
  lemonEty:etyTarget exlex:frater-la ;
  rdfs:label "Etymology Link" .
```

Finally, RDF generation must account for the scale of the Wiktionary-derived data. The construction of triples itself is relatively inexpensive and, if the pipeline were purely CPU-bound, it would run substantially faster. In practice, however, the enrichment stage dominates runtime, since it depends on external network requests to resolve Wiktionary URLs and query the LiLa linking service. To keep the pipeline scalable, the graph is not stored in memory as a whole; instead, each enriched etymology chain is immediately serialized to an `.nt` file. This streaming approach prevents excessive RAM usage and is therefore essential when processing large volumes of data. N-Triples is especially appropriate for this purpose because it supports incremental, line-based serialization. Finally, for human inspection, the resulting `.nt` file is converted into Turtle (`.ttl`), a more readable RDF format.

3.5 Data publication

The resulting dataset has been published as a Linked Data resource within the LiLa graph and is accessible online through the LiLa infrastructure at https://lila-erc.eu/data/lexicalResources/englishWiktionaryLatinEtymologies/Lexicon. Entitled *English Wiktionary Latin Etymologies*, it contains 9,684 lexical entries corresponding to Latin lemmas with curated etymologies extracted from the English edition of Wiktionary. Each entry is represented as an `ontolex:LexicalEntry` linked to its etymological representation through the lemonEty ontology. The lexicon metadata specifies the provenance of the data, indicating that the etymologies were retrieved from the raw Wiktextract dump made available at https://kaikki.org/dictionary/rawdata.html and processed into RDF by our pipeline. The dataset is released under the same license as the underlying Wiktionary data, Creative Commons Attribution-ShareAlike (CC BY-SA), and is integrated into the LiLa ecosystem[3], which allows it to interoperate with existing Latin linguistic resources already linked within the knowledge base.

3.6 Reproducibility

The corpus generation pipeline is fully reproducible. It begins with the raw Wiktextract dump of the English Wiktionary available at https://kaikki.org/dictionary/rawdata.html. From this source, etymology templates are extracted and converted into structured etymology chains. These chains are then curated, enriched and, finally, serialized into RDF triples. The code implementing this pipeline can be retrieved from the GitHub repository associated with this project, https://github.com/CIRCSE/englishWiktionaryLatinEtymologies , in which the dataset in both .nt and .ttl can be found

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

as well. This allows the corpus generation process to be reproduced from the original Wiktextract data. That is, given a Wiktextract dump, the pipeline consistently returns the same RDF output. The only exception is the entity linking step, where manual disambiguation may be required in cases of lexical ambiguity, as discussed next in the Limitations section.

4 Limitations

The dataset presented in this work is derived from the English version of Wiktionary and therefore inherits some of its properties. Multiple entries may exist for different grammatical forms of the same lexical item (e.g., nominative *animal* and its genitive *animalis*) or for distinct uses across parts of speech (e.g., *Februārius* as both a proper noun and an adjective). Wiktionary provides valuable etymological information, but some degree of noise is inherent to the source data.

A further limitation concerns the alignment with external resources, such as the Lemma Bank of LiLa. In some cases, an entity linking service yields several candidates, as ambiguity cannot be resolved through surface form matching alone. As discussed in the RDF Serialization section, the current pipeline always selects the lemma with the lowest value for the lemma ID, but accurate disambiguation requires manual inspection. In our data, 381 lemmas were identified as having multiple candidate matches. While this does not affect the structural validity of the etymological chains, it impacts the precision of the alignment with LiLa and sets direction for future work.

For example, for the Latin form *os*, the LiLa Text Linker yields the candidates 115327 (“mouth, face, opening”), 115330 (“bone”) and 68537 (no meaning available). *Os* has two etymology chains:

`[(h₃éh₁os,ine-pro),(ōs,itc-pro),(os,la)]`

`[(h₃ésth₁, ine-pro), (os, la)]`

In Wiktionary, we can see that the etymology chain tracing back to Proto-Indo-European *h₃éh₁os* corresponds to the sense “mouth” and aligns with LiLa lemma 115327, whereas the chain derived from *h₃ésth₁* corresponds to “bone” and matches LiLa lemma 115330. This illustrates that selecting the correct lemma requires interpreting the meaning associated with each etymology and aligning it with the corresponding entry in the external resource.

Finally, the curation stage prioritizes precision over recall. While this reduces noise and increases the overall reliability of the dataset, it may also exclude valid etymological relations. As a result, there may be false negatives that have been excluded from the dataset, which should therefore not be considered an exhaustive collection of all Latin etymologies present in Wiktionary.

5 Conclusion

In this paper, we presented a curated resource of Latin etymological chains automatically extracted from the English edition of Wiktionary and represented as RDF triples using the OntoLex-lemon model and its lemonEty extension. The resulting dataset contains validated etymology chains for 9,684 lemmas and has been integrated into the LiLa Knowledge Base of Linguistic Resources for Latin, enabling interoperability with existing linked linguistic resources.

The pipeline used to generate the corpus is fully reproducible and can be adapted to extract etymological data for other languages available in Wiktionary. We hope that the resulting resource will support future work in Historical Linguistics, computational etymology, and linked lexical data, as well as facilitate the integration of etymological information into broader linguistic knowledge graphs.

References

Pierpaolo Basile and 1 others. 2022. A new time-sensitive model of linguistic knowledge for digital humanities. In *Proceedings of the Workshop on Language Technology for Digital Humanities*.

Tim Berners-Lee. 2006. Linked data.

Anas Fahad Khan. 2018. Towards the representation of etymological data on the semantic web. *Information*, 9(11):288.

Francesco Mambrini and Marco Passarotti. 2020. Representing etymology in the lila knowledge base of linguistic resources for latin. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 20–28, Marseille, France. European Language Resources Association.

John P. McCrae and 1 others. 2017. The ontolex-lemon model: Development and applications. *Electronic Lexicography in the 21st Century*.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic

network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Ester Pantaleo, Vito Walter Anelli, Tommaso Di Noia, and Gilles Sérasset. 2017. Etytree: A graphical and interactive etymology dictionary based on wiktionary. In *Proceedings of the WWW Companion*.

Marco Passarotti, Francesco Mambrini, and 1 others. 2024. The lila text linker: Linking latin texts to the lila knowledge base. In *Proceedings of the Workshop on Linked Data in Linguistics (LDL 2024)*.

Gilles Sérasset. 2014. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1317–1325, Marseille, France. European Language Resources Association.