Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin

Francesco Mambrini, Marco Passarotti
CIRCSE Research Centre, Università Cattolica del Sacro Cuore
Largo Gemelli, 1 - 20123 Milan (Italy)
{francesco.mambrini, marco.passarotti}@unicatt.it

Abstract

In this paper we describe the process of inclusion of etymological information in a knowledge base of interoperable Latin linguistic resources developed in the context of the *LiLa: Linking Latin* project. Interoperability is obtained by applying the Linked Open Data principles. Particularly, an extensive collection of Latin lemmas is used to link the (distributed) resources. For the etymology, we rely on the Ontolex-lemon ontology and the lemonEty extension to model the information, while the source data are taken from a recent etymological dictionary of Latin. As a result, the collection of lemmas LiLa is built around now includes 1,465 Proto-Italic and 1,393 Proto-Indo-European reconstructed forms that are used to explain the history of 1,400 Latin words. We discuss the motivation, methodology and modeling strategies of the work, as well as its possible applications and potential future developments.

Keywords: etymology, linked open data, Latin

1. Introduction

Latin is the most widely attested member of the Italic branch of the Indo-European family, which also includes other cognate languages (such as for instance Oscan, Umbrian and Faliscan) spoken in central and southern Italy before the Roman domination. As the language of Rome, whose authority and influence extended over the Mediterranean as well as a large portion of continental Europe and of the Near East for many centuries, Latin played a role in the cultural and linguistic history of the world that is hard to overestimate. Moreover, as the direct ancestor of the Romance family, several languages of Europe like Spanish, Portuguese, French, Italian and Romanian trace their roots directly to it. As a consequence, a great part of the vocabulary of many modern languages is derived, through inheritance or borrowing, from Latin.

In the present days, large corpora of Latin texts for several million words, belonging to different genres and produced in the span of many centuries, are publicly available on the web.[1] In addition to texts, the internet provides also an extensive selection of digitized dictionaries, including etymological lexica (Mambrini and Passarotti, 2019, 72-3 for an overview). While these resources can be browsed, read and queried from separate interfaces, interaction between them is extremely limited.

Indeed, etymological studies are a very good example of how the lack of interoperability between digital resources imposes limitations to users. Researchers and students of historical linguistics would greatly benefit from the capability to interrogate simultaneously all the dictionaries that discuss the etymology, meaning or synonyms of words, together with corpora that document all the attestations of any given lexical item. However, this experience is precluded by the limits of the publication model currently used for lexica and corpora, which relegates them in the condition of isolated silos.

The adoption of the Linked Open Data (LOD) paradigm for linguistic resources can greatly improve the situation for historical linguistics of Latin. Defined by Berners-Lee with the goal of shifting from a web of document to a web of interconnected data (Berners-Lee, 2006), the LOD principles prescribe, among other things, to use Uniform Resource Identifiers (URIs) as names, preferably in the form of HTTP URLs that can be looked out on the web, and to include links to other URIs so as to provide context for the published data. The advantage of the model for linguistic resources is evident, as in a web of data it becomes "possible to follow links between existing resources to find other, related data and exploit network effects" (Chiarcos et al., 2013, iii). Not by chance, across the last years the research community dealing with the creation and distribution of linguistic resources has been working extensively to build the so-called Linguistic Linked Open Data cloud (LLOD),[2] a collaborative effort pursued by several members of the Open Linguistics Working Group,[3] with the goal of developing a Linked Open Data (sub-)cloud of linguistic resources as part of the wider Semantic Web (McCrae et al., 2016).

In this context, the Ontology-Lexica Community Group has been particularly active in the effort to provide models for the representation of lexica as LOD. The main result of the enterprise is the publication of the Ontolex-lemon model, now a de facto standard for the representation of lexical resources (McCrae et al., 2017).[4]

Ontolex is built around a core module, whose primary element is the Lexical Entry; this class includes all the relevant elements of the lexicon, such as words, multi-word expressions or morphemes like affixes. Lexical entries are connected to forms that represent the grammatical realiza-

[1] To give an idea, on March 22, 2019, the (meta-)repository of Latin corpora *Corpus Corporum* (`http://www.mlat.uzh.ch/MLS/`) passed the total of 160 million words with its latest update.

[2] `http://linguistic-lod.org/llod-cloud`.

[3] `https://linguistics.okfn.org/index.html`.

[4] `https://www.w3.org/2016/05/ontolex/`.

Proceedings of the Globalex Workshop on Linked Lexicography , pages 20–28
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

tion of the lexical item; one of them can be identified as the canonical "dictionary form", or lemma. From the standpoint of meaning, entries can be linked to concepts in ontologies either directly (through a denotative link) or via a "lexical sense" that reifies the relation between an entity from an ontology (e.g. a concept from DBpedia)[5] and a lexical entry.

The Ontolex-lemon model has been extended to account for a number of linguistic properties of the lexicon, like translation (Gracia et al., 2014) and lexicographic metadata.[6] Most recently, Khan (2018a) proposed an extension of Ontolex, called lemonEty, designed to represent also etymological information linked to lexical entries. The extended Ontolex-lemon model is therefore suitable to represent complex lexicographic information, including etymology, in the Semantic Web; this, in turn, is a step towards interoperability between resources, which, as we saw, is a fundamental *desideratum* for students and researchers in (historical) linguistics.

Other approaches to the task of modeling etymological lexical resources using LOD principles include the endeavor to represent the *Dictionnaire étymologique de l'ancien français* (DEAF) (Städtler et al., 2014) using OntoLex-Lemon (Tittel and Chiarcos, 2018) and the LOD representation, again using Lemon, of the *Tower of Babel (Starling)*, a major etymological database featuring short- and long-range etymological relations (Abromeit et al., 2016).

As for Latin, Bon and Nowak (2013) show how intrinsic wiki concepts, such as namespaces, templates and property-value pairs can be used for linking Medieval Latin dictionaries. The same authors are also among the developers of *medialatinitas.eu*,[7] a Web application that integrates dictionaries, corpora and encyclopaedic resources for Latin in a user-friendly interface, although it does not provide any explicit (and reusable) link between the resources (Nowak and Bon, 2015).

The idea of using the LOD paradigm to integrate not only lexical resources, but also textual corpora and Natural Language Processing (NLP) tools for Latin in the Semantic Web is the guiding principle of the project *LiLa: Linking Latin* (henceforth, LiLa). This paper reports on a large-scale experiment on including the information from a recent etymological dictionary of Latin and Italic languages into the Ontolex-based lexical knowledge base of Latin canonical forms of LiLa. Section 2. summarizes the aims and the current status of LiLa. Section 3. describes the treatment of etymology in the LiLa knowledge base. Particularly, 3.1. presents the source of our etymological data; 3.2. provides more details on the lemonEty ontology that was adopted for the experiment; 3.3. discusses the representation of etymologies as scientific propositions, and 3.4. describes how we integrated the etymological information into the LiLa architecture. Section 4. reports an example of how we can make the etymologies interact with the rest of the linguistic information in LiLa. Finally, Section 5. concludes the paper and outlines directions for future work.

[5] https://wiki.dbpedia.org/.

[6] https://www.w3.org/2019/09/lexicog/.

[7] https://medialatinitas.eu/.

2. LiLa: Linked Open Data for Latin resources

The ERC-funded LiLa project (2018-2023) intends to use the LOD paradigm to build a knowledge base of linguistic resources for Latin, i.e. a collection of several (distributed) data sets described using the same vocabulary of knowledge description and linked together.[8] Ultimately, the goal of LiLa is to exploit the wealth of linguistic resources and NLP tools for Latin developed thus far to the best, in order to bridge the gap between raw language data, NLP and knowledge description (Declerck et al., 2012).

The approach adopted by LiLa rests on two principles. Our initial assumption is that lexicon is the level where interoperability between linguistic resources can be achieved, as texts are made of occurrences of words, lexica and dictionaries describe properties of words, and NLP tools process words. But, in particular for a richly inflected language like Latin, the level of lemma is considered the ideal interface between the different types of resources we intend to link. Lemmatization, defined as the task to reduce the inflected forms of a word to one of them conventionally chosen to be the canonical form (e.g. the first person singular of indicative for verbs), is a layer of annotation common to different kinds or resources. Dictionaries tend to index lexical entries using lemmas. Thesauri organize the lexicon by collecting all related entries, and use lemmas to index them. Digital libraries use lemmas to enable lexical search in corpora. In NLP, lemmatization is also included in many pipelines of annotation.

The core of the LiLa knowledge base is built around a comprehensive collection of Latin forms that can be used as lemmas in lexical or textual resources. As we said, in the Ontolex-lemon model the traditional notion of "lemma" is expressed by the "canonical form" property that links a lexical entry to one (and not more than one) form. Therefore, by modeling our collection of lemmas as Ontolex's forms that are potentially used as canonical forms of lexical entries, we ensure compatibility with any other resource that adopts that ontology. As Ontolex forms are licensed to have multiple written representations, the model is very apt to express any orthographic variation and non-canonical spelling of words, which is particularly important for a language like Latin with more than 2,300 years of written attestation.

The list of lemmas included in LiLa was populated from the comprehensive database of the Latin morphological analyzer Lemlat (Passarotti et al., 2017). Lemlat's database reconciles three reference dictionaries for Classical Latin (Gradenwitz, 1904; Georges and Georges, 1913 1918; Glare, 1982), the entire Onomasticon from Forcellini's *Lexicon Totius Latinitatis* (Budassi and Passarotti, 2016), and the Medieval Latin *Glossarium Mediae et Infimae Latinitatis* by du Cange et al. (1883 1887), for a total of over 150,000 lemmas (Cecchini et al., 2018).

Currently, LiLa includes 190,237 lemmas.[9] The relevant

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

[9] The current total number of lemmas in the LiLa collection is higher than in Lemlat, because LiLa includes also a set of lemmas for deadjectival adverbs and present, future and perfect participles

morphological properties of them (part of speech, gender, inflection type) are described using a specific ontology that we intend to align with OLiA (Chiarcos and Sukhareva, 2015). This collection is what the etymological information is linked to and that ultimately serves as a connection point with the other linguistic resources on Latin.

The portion of LiLa that is based on the list taken from the aforementioned three dictionaries of Classical Latin was also enriched with information on word formation derived from the lexicon of the project *Word Formation Latin* (WFL) (Litta et al., 2016).[10] In LiLa, all the lemmas analyzed in WFL are connected to the derivational morphemes (prefixes and affixes) and the lexical bases that can be isolated in them. Thus, it is possible to browse, for instance, all the canonical forms where the prefix *ad-* is used,[11] or the 12 lemmas that have the same lexical base as the noun *rosa* "rose" (Litta et al., 2019).[12]

3. Etymologies in LiLa

3.1. Data

An etymological dictionary is a lexicon that aims to reconstruct the history of each entry, rather than focusing on aspects of meaning or usage. In this context, etymology is generally intended as the task of documenting the origins of a given lexical item and trace back its transfers across different languages, be it by borrowing (even from genetically unrelated tongues), or in a direct hereditary relation from an ancestor to the target language. In the case of the earliest attested Indo-European languages like Latin, particular stress is put on the latter phenomenon. Historical linguists attempt, whenever it is possible, to investigate the most remote origin, form and meaning of a word in the Proto-Indo-European (PIE) phase, based on the comparative study of the evidence offered by the cognate languages, and/or in the intermediate (also reconstructed) ancestor of a sub-family (like the Proto-Italic, henceforth PIt, for the Italic family). Less frequent, but obviously not less interesting, is the case of words that don't appear to have a plausible Indo-European etymology and are (often, very tentatively) explained as loans from non-Indo-European languages.

The etymological information that we connect to the LiLa knowledge base is taken from the most recent *Etymological Dictionary of Latin and the other Italic Languages* (de Vaan, 2008). The content of the dictionary itself is copyrighted by the publisher; however, the owners (Brill ed.) have clarified to us *per litteras* that information about the reconstructed PIt and PIE forms and their connection to the Latin words can be used, provided that explicit attribution to the author and the publication is given.

The dictionary contains 1,874 entries, which, as it is customary for etymological lexica, do not cover the whole Latin vocabulary. Words created by regular derivation processes internal to a language (e.g. by derivational morphemes) are generally grouped together under whatever word is identified as the most interesting for etymological purposes.[13] So, for instance, the nouns *aedicula* "small house" (formed with the diminutive suffix *-cul*) and *aedilis* "aedile" (a magistrate for public buildings, formed with the suffix *-il*) do not have an etry for themselves in the dictionary, but are instead listed among the derivatives of *aedes* "dwelling-place, temple". Also, the entries are limited to the words that belong to the inherited lexicon of Latin: the loan words (mostly from Ancient Greek), which are especially frequent in the domains of grammar, science and philosophy, are not treated.

In the dictionary by de Vaan (2008), each entry follows a defined structure, in which five layers can be distinguished. The first level provides the lemma, a translation, a minimal historical contextualization (such as the first attestation), some relevant morphological information (part of speech, gender and inflection type), and a series of Latin cognate words (like *aedicula* and *aedilis* for *aedes*). The following sections list the PIt and PIE reconstructed ancestors, together with a set of cognate words attested (or postulated) in the related languages. Finally, the last two paragraphs contain a lengthier discussion of the history of the word and a bibliography.

As per the agreement with the publisher, we modeled only the information about the PIt and PIE reconstructed ancestors in the second level of the structure just described. The goal is to introduce such ancestors into LiLa, by linking them, according to the chosen ontology, to the relevant Latin lemmas of the LiLa's collection.

Of the ca. 1,900 entries in de Vaan (2008), we identified 1,466 that explicitly list a PIE and/or a PIt reconstructed etymology in the paragraph that we targeted for extraction. Another 25 of them belong to an Italic language (mostly Oscan or Umbrian) and are therefore not linkable to LiLa. A final group of 50 entries that we could not properly link are those that discuss the etymology of derivational morphemes; although, as said, LiLa does provide information on prefixes and suffixes, these morphemes are still not represented as lexical entries in our knowledge base, thus making it impossible to use a Ontolex-based model to describe their etymology.

In total, we identified a pool of 1,391 entries from de Vaan (2008) for which etymological information could be linked to a LiLa lemma.

3.2. The model

The Ontolex-lemon Etymological Extension or lemonEty (Khan, 2018a; Khan, 2018b) extends the Ontolex core by introducing a number of classes and properties to encode etymological information about lexical entries.

The first new class is the Etymology itself. The class reifies the whole process of etymological reconstruction as scientific hypothesis; the main advantage of this approach is that it allows to make statements about the etymology itself, such as the attribution to scholars, bibliographical ref-

that were automatically built from the Lemlat database. For more details, see Mambrini and Passarotti (2019).

[10] `http://wfl.marginalia.it/`.

[11] `https://lila-erc.eu/data/id/prefix/5`.

[12] `https://lila-erc.eu/data/id/base/3079`.

[13] According to de Vaan (2008, 10), the word chosen for the entry in the dictionary "represents the derivationally most opaque member of a Latin word family". We take this to mean the word whose derivation cannot be explained (or is explained less easily) with the regular Latin word-formation rules.

erences, or belief values, so that the model can theoretically include also discarded hypotheses that are considered not plausible by specialists (see below, Section 3.3.).

Etymologies group together a series of related lexical entries, one of which (identified by the "lemma" of the entry in the etymological dictionary) is the target whose history must be explained. Any lexical item that is introduced only to describe the history of a word and, as a rule, does not belong to the lexicon of a given language, is a member of the Etymon class, a subclass of Ontolex's Lexical Entry. The subclass serves the purpose of maintaining a distinction between the proper lexical entries of a given language and those words (from an ancestor or any other languages or language phase) that are introduced only for the etymological purposes.

Although the hypotheses concerning the origins and histories of words can be quite complex, and may involve transfers of meanings or restructuring of forms, etymologies can in general be conceptualized as sequences of steps from an earlier linguistic stage to a subsequent phase, until the target word is satisfactory explained. Thus, for instance, Lat. *lupus* "wolf" is explained by de Vaan (2008, 353) by posing a passage from PIE $*ulk^wo$- to PIt $*luk^wo$- by metathesis, and from the latter to Latin (possibly, via a loan from Sabellic).[14]

The lemonEty extension allows to model such sequences of stages with the help of the class Etymology Link. An Etymology Link reifies the etymological relations between a source (i.e. an expression postulated as the origin of the relation, such as a word in the ancestor language) and a target. In the example quoted above, the etymology of *lupus* implies the existence of three etymology links: PIE $>$ PIt ($>$ Sabellic) $>$ Latin. The links can then be further specified by defining the type of relations that they imply; in the example, the links between PIE and PIt and from PIt to any Italic language imply inheritance, while the one between Sabellic and Latin is a borrowing. The "sub-source" property can also be attached to the link, in order to narrow some specific semantic or morphological properties of the source word that are relevant for the process. So, for instance, a sublink can be used to specify that Italian *lupo* "wolf" is derived from the accusative form (*lupu(m)*) of Latin *lupus*.

Figure 1 reproduces the proposed etymology for *lupus*, as represented in LiLa,[15] with the links from the reconstructed PIE word to the reconstructed PIt and from PIt to Latin.[16]

[14]Metathesis is a process of transposition of syllables or phonemes that is fairly common in the history of words: see for instance Italian *coccodrillo* or Spanish *cocodrilo* "crocodile" from Latin *crocodilus*. Note that in historical linguistics the asterisk is the conventional mark for reconstructed forms, i.e. those forms that, although not positively documented, are postulated by applying the comparative method.

[15]`https://lila-erc.eu/data/lexicalResources/BrillEDL/id/etymology/178`.

[16]Since, as we said, we decided to limit our work to PIE and PIt etymons, the etymological representation in LiLa at present skips the passage from Sabellic to Latin.

3.3. Etymologies as scientific propositions

An important feature of the lemonEty ontology is that it allows to represent etymologies as a set of propositions about the history of words, which can be properly attributed and described with all properties pertaining to scientific discourse.

The approach that we adopted to model etymologies as scholarly output is based on the CIDOC Conceptual Reference Model (CRM) (Doerr, 2003), a widely adopted formal ontology used for heterogeneous cultural heritage information. In terms of the CIDOC-CRM, etymologies can be considered instances of the class "E89 Propositional Object", which encompasses the "sets of propositions about real or imaginary things and that are documented as single units or serve as topic of discourse";[17] examples of E89 include Maxwell's Equations or Anselm's ontological argument. The property P70 ("documents") can be used to link any "E31 Document" to any entity of the CIDOC CRM.[18] Therefore, the statement expressing that de Vaan's dictionary (an instance sof E31) documents (via the P70 property) an etymology like the one represented in Figure 1 (E89) is a suitable way to encode the bibliographical attribution. This modelization is represented in Figure 2.

In our first experiment, we limited ourselves to this very simple set of statements. However, as PIE reconstruction is a very speculative field, the model can be enhanced to capture more nuances of the sometimes complex domain of etymological argumentation. In the following paragraphs, we propose a possible modelization that, although not (yet) implemented in LiLa, may be advisable in order to make the information that we derived from de Vaan (2008) more interoperable with other etymologies that are published (or that may be published) on the web.

In his discussion on the history of *lupus*, de Vaan (2008, 353) mentions an alternative hypothesis to the one adopted in LiLa (represented in Figure 1), which he considers less persuasive. According to this alternative reconstruction, the word may originate from PIE $*ulp$-/$*lup$- "marten" (see Latin *volpes* "fox"), with a semantic shift from the original sense to the one of wolf.

While, as we saw, lemonEty is equiped to express the semantic change from PIE to Latin, we can apply the CRM$_{inf}$ (Argumentation Model) extension of the CIDOC CRM (Stead et al., 2019) to represent the whole process of argumentation that is reflected in the entry of the etymological dictionary.[19] An Etymology, with its attached Etymology Links and Etymons, can be considered an instantiation of a "I4 Proposition Set" as defined by the Argumentation Model. These propositions are then associated to a belief value (for instance, true or false) in instances of the class "I2 Belief".[20] The CRM$_{inf}$ can thus be used to express the

[17]`http://www.cidoc-crm.org/html/5.0.4/cidoc-crm.html#E89`.

[18]The 'E31 Document' is the class that "comprises identifiable immaterial items that make propositions about reality" (`http://www.cidoc-crm.org/html/5.0.4/cidoc-crm.html#E31`).

[19]`http://new.cidoc-crm.org/crminf/`.

[20]The class I2 "comprises the notion that the associated I4 Proposition Set is held to have a particular I6 Belief Value by a

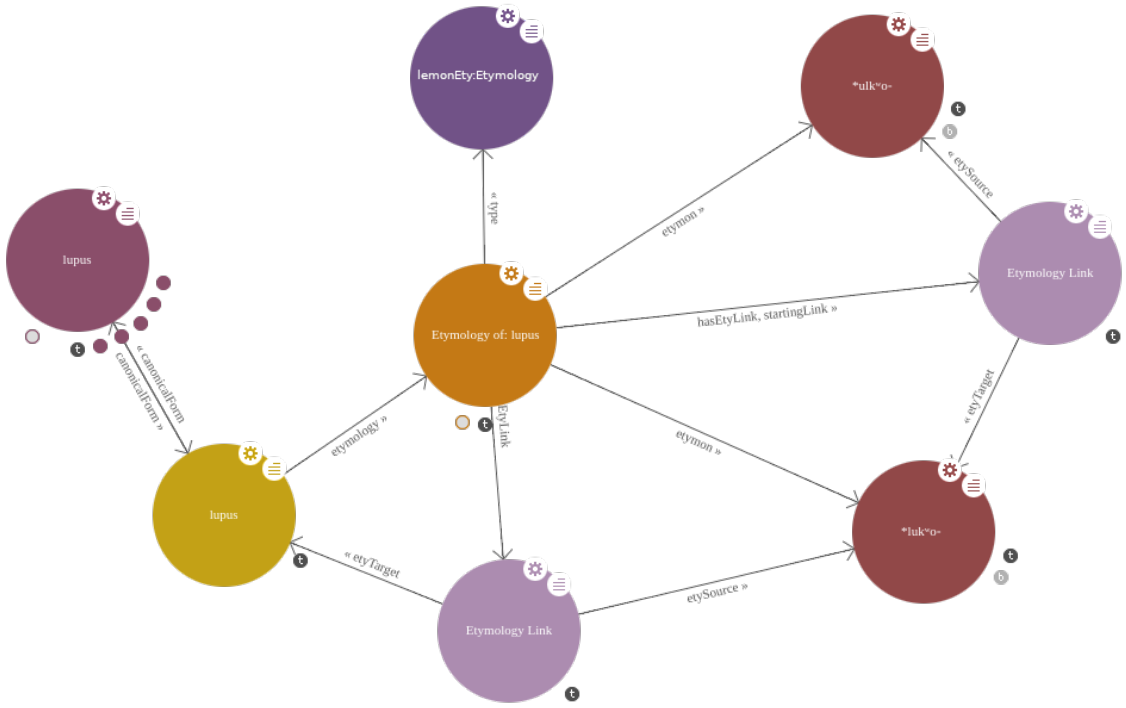

Figure 1: The etymology of *lupus* in LiLa according to the lemonEty model.

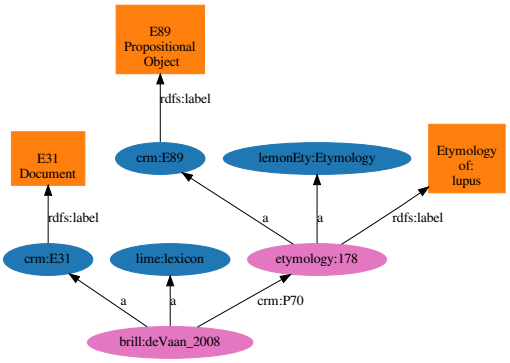

Figure 2: The etymology of *lupus* as an "E89 Propositional Object".

fact that the content of each I4 is considered true or false. Figure 3 illustrates a schematic representation of de Vaan's etymological argumentation about *lupus* according to this model. The whole process of discussion is represented as an instance of the "I1 Argumentation" class.[21] The conclusions are represented by two beliefs ("J2 concluded that"); on the one hand, the etymology represented in Figure 1 (and not reported in Figure 3) is held to be true, while the second belief is that an alternative explanation (shown here with a single etymology link to PIE $*ulp$-/$*lup$-) is considered less plausible.

For the sake of simplification, Figure 3 adopts a black-and-white model of belief values, where only "True" and "False" are distinguished. de Vaan (2008, 353) uses a much more nuanced language: the accepted explanation is "conceivable", while the alternative entails assumptions that "would require further special pleading". It should be noted that even these assumptions can be encoded using the model suggested here. In Figure 3, the semantic shift required is encoded in the lexical senses attached to the two lexical entries connected via the etymology link; the source is the PIE etymon with the postulated sense of "marten", while the Latin target refers to a different animal (the wolf). The main problem of this etymology, according to de Vaan, is to explain the fact that the root is also continued by Latin *volpes* "fox"; although not shown in Figure 3, it is clear

particular E39 Actor. This can be understood as the period of time that an individual or group holds a particular set of propositions to be true, false or somewhere in between" (Stead et al., 2019, 10).

[21] An I1 represets "the activity of making honest inferences or observations. An honest inference or observation is one in which the E39 Actor carrying out the I1 Argumentation justifies and believes that the I6 Belief Value associated with resulting I2 Belief about the I4 Proposition Set is the correct value at the time that the activity was undertaken" (Stead et al., 2019, 10).

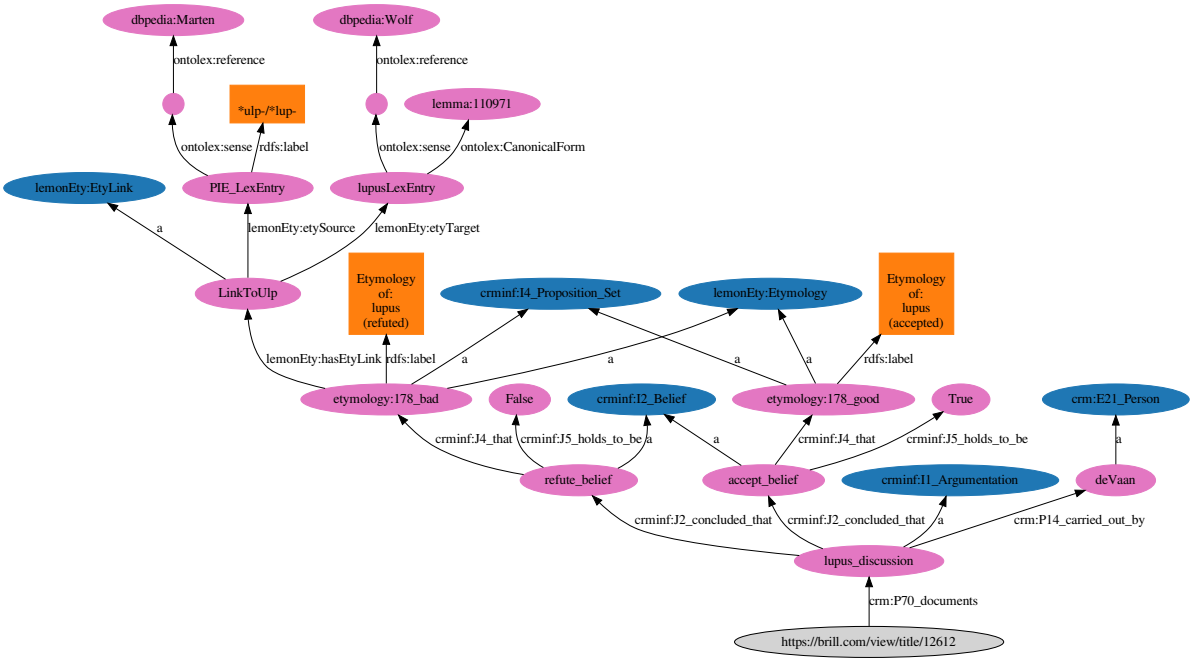

Figure 3: Using CRM$_{inf}$ to model the argumentation about the etymology of *lupus*.

from our previous discussion that the lemonEty model is capable of representing the etymologies of these two words converging to this PIE etymon.

3.4. The linking process

A scrutiny of the 1,391 entries taken from de Vaan (2008) led us to include 2,858 instances of Etymon in LiLa, 1,465 for PIt and 1,393 for PIE. These entries are now grouped in 1,434 etymologies and linked by 2,648 etymology links.

1,400 lexical entries from the Brill dictionary are now connected to a lemma from the LiLa collection. Of these, 1,383 are also linked to an Etymology, while 17 are cognates of other words that share every trait with them (including obviously the history) but the part of speech, like for instance *supra* "over", which is assigned both part of speech adverb and adposition.[22]

In the process of linking, we encountered several cases where the lemma of the dictionary entry matched more than one lemma of the LiLa collection. For instance, the string "pullus" could be matched to three different lemmas of LiLa: one noun ("foal"), and two adjectives, meaning "pure" and "dark-colored" (the latter being the correct lemma for the entry in de Vaan).[23] In all but 13 cases, a manual disambiguation allowed us to identify the correct candidate; most often, as in the case of *pullus*, the information on the part of speech, the inflection class or the derivational morphology attached to each lemma was sufficient to disambiguate. The other 13 cases either involve errors in the morphological annotation or reflect a greater ambiguity that requires further study.

The connections between the etymons, etymology links and lemmas can be queried using the SPARQL endpoint of LiLa.[24]

4. Using etymologies as linked data

Once that the etymologies are linked with the LiLa lemmas, it becomes possible to cross the information on PIE and PIt derivation with the other resources represented in our knowledge base.

One possible example of a meaningful connection that can be explored is that between etymology and word formation. As we saw, the entries in de Vaan (2008) cover only a portion of the Latin lexicon; some words are listed as cognates and derivatives of the main entry, while many more secondary formations, especially of late attestation, are not mentioned at all. The entry dedicated to *clārus* "loud, bright" in de Vaan (2008, 117-118), for instance, reports six words as Latin cognates, but some other like *clarificatio* "glorification", attested in Ecclesiastical Latin, are not listed. According to the index of Latin forms in de Vaan (2008, 725-765), the dictionary discusses 9,439 Latin words (including affixes).

The information about the derivational morphemes in LiLa may help retrieving the other derivative words that are not explicitly mentioned by de Vaan. Following the model

[22] The Ontolex-lemon model requires that in such cases as many lexical entries are created as are the relevant assignments of part of speech.

[23] https://lila-erc.eu/data/id/lemma/120692.

[24] https://lila-erc.eu/sparql.

of Construction Morphology (Booij, 2010), used to represent in LiLa the derivational information provided by WFL (Litta et al., 2019), 36,318 lemmas in the lexical collection of LiLa (corresponding to the section of the analyzer Lemlat optimized for Classical Latin) are linked to the prefixes, the suffixes and the lexical bases that can be distinguished in their internal structure. Thus, the noun *clarificatio* mentioned above is connected to two lexical bases (the one shared with *clarus* and the one shared with *facio* "to make") and the deverbative suffix *-(t)io(n)*.

Lexical bases provide a suitable starting point to investigate the links. As a rule, words that share a lexical base with a lemma of an entry in de Vaan (2008) also share the same etymology. A SPARQL query over LiLa's endpoint returns 1,200 bases out of 3,858 (31.10%) that are linked to at least one lemma of a lexical entry connected to an etymology.

Although these 1,200 bases cover less than a third of the total in LiLa, they link 23,292 lemmas (64% of the lemmas attached to a base). In fact, on average, bases that are connected to a word linked to an etymology group a significantly larger numbers of lemmas (21.01) than those with no link to etymologies (5.36). This may be due to several concurring factors. Some words of PIE origin (such as *facio* "to make", *fero* "to bring" or the numeral *tres* "three") are extremely productive.[25] On the other hand, many loan words, which, as we said, are not discussed by de Vaan (2008) and thus have no etymology link, are usually technical terms that gave origin (if at all) to very few derivatives.

The connection with the lexical bases in the LiLa knowledge base allows us to supplement the list of the ca. 9,400 derivatives with a number of new units ranging from 13,853 new units (assuming all the words in the index of de Vaan are in the results) to a maximum of 23,292 (if no words in the index are in the results). In either case, this represents a significant increase in the coverage of the Latin lexicon.

5. Conclusions and future work

By adopting the Ontolex-lemon model and the lemonEty expansion, and building on the LiLa's original assumption of linking through lemmatization, we were able to include a basic set of etymological connections to our knowledge base of Latin canonical forms. Namely, we introduced a set of etymologies, defined as scientific hypotheses about the inheritance links between Latin words and the reconstructed forms in the PIt and PIE languages. It is now possible to follow the links from the etymologies to the lemmas and, from there, to all the other resources connected to each canonical form.

The potential applications of the (meta)data we created are several and, most importantly, transcend the limits of Latin linguistics. Latin is in fact just one of the many languages that trace their root to PIE. To go back to the example of *lupus* and to name but a few random modern languages, words as different as English *wolf*, Irish *olc*, Czech *vlk*, Albanian *ujk*, Greek *lýkos*, Hindi *vŕk* and Persian *gorg* all originate from the same reconstructed PIE word. Potentially, all lexical databases for Indo-European languages could have etymological links pointing to the same PIE etymon.

[25] In LiLa, the bases linked to these three lemmas count 688, 367 and 36 lemmas respectively.

In most cases, the precise reconstruction of the form and meaning of a PIE etymon will be extremely controversial. Although this field of research is very speculative, and strong disagreement and incompatible hypotheses are often the rule rather than the exception, we have shown that the ontologies available are capable of modeling, at least broadly, the terms of the scholarly debate and to capture the arguments.

While the information encoded in LiLa is already rich, several directions for future improvement can be outlined. On the one hand, the Latin derivatives of each entry mentioned by de Vaan (2008) and included in the index of 9,439 words mentioned above can be explicitly linked to the main lexical entries as cognates. Also, the etymology of some selected affixes of PIE or PIt origin can be attached to the relevant morphemes by using the Ontolex-lemon model and lemonEty.

Finally, we intend to link the Latin WordNet (LWN) (Minozzi, 2010) to our lemmas, but also to increase its coverage (Franzini et al., 2019). The connection with LWN would allow us to expand the etymology links to trace the sub-links to the senses of Latin lexical entries and the meaning of the PIt and PIE as reconstructed from the comparative evidence. The process of mapping the semantics of etymons would thus produce a similar output to that visualized in Figure 3 for the discarded etymology of *lupus*, with WordNet used as a reference ontology instead of DBpedia.

6. Acknowledgements

We are grateful to Prof. Michiel de Vaan for proposing us to use his work in our project, and to Brill Ed. for granting us the permission to link some of the information contained in the Latin etymological dictionary to LiLa.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme - Grant Agreement No 769994.

7. Bibliographical References

Abromeit, F., Chiarcos, C., Fäth, C., and Ionov, M. (2016). Linking the tower of babel: modelling a massive set of etymological dictionaries as rdf. In *Proceedings of the 5th Workshop on Linked Data in Linguistics (LDL-2016): Managing, Building and Using Linked Language Resources*, pages 11–19, Portoroz, Slovenia. European Language Resources Association (ELRA).

Berners-Lee, T. (2006). Linked data. `https://www.w3.org/DesignIssues/LinkedData.html`. Accessed February 13, 2020.

Bon, B. and Nowak, K. (2013). Wikilexicographica. linking medieval latin dictionaries with semantic mediawiki. In *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013*, pages 407–420, Ljubljana and Tallinn. Trojina, Institute for Applied Slovene Studies and Eesti Keele Instituut.

Booij, G. (2010). Construction morphology. *Language and linguistics compass*, 4(7):543–555.

Budassi, M. and Passarotti, M. (2016). Nomen omen. Enhancing the Latin morphological analyser Lemlat with

an onomasticon. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 90–94, Berlin, Germany. Association for Computational Linguistics.

Cecchini, F., Passarotti, M., Ruffolo, P., Testori, M., Draetta, L., Fieromonte, M., Liano, A., Marini, C., and Piantanida, G. (2018). Enhancing the latin morphological analyser lemlat with a medieval latin glossary. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018). 10-12 December 2018, Torino*, pages 87–92, Torino. aAccademia University Press.

Chiarcos, C. and Sukhareva, M. (2015). OLiA - Ontologies of Linguistic Annotation. *Semantic Web Journal*, 6(4):379–386.

Chiarcos, C., Cimiano, P., Declerck, T., and McCrae, J. P. (2013). Linguistic linked open data (llod). Introduction and overview. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages i–xi, Pisa, Italy. Association for Computational Linguistics.

de Vaan, M. (2008). *Etymological Dictionary of Latin: and the other Italic Languages*. Brill, Amsterdam.

Declerck, T., Lendvai, P., Mörth, K., Budin, G., and Váradi, T. (2012). Towards linked language data for digital humanities. In *Linked Data in Linguistics*, pages 109–116. Springer, Berlin.

Doerr, M. (2003). The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75–75.

du Cange, C. d., Bénédictins de Saint-Maur, Carpentier, P., Henschel, L., and Favre, L. (1883–1887). *Glossarium mediae et infimae latinitatis*. Favre, Niort, France.

Franzini, G., Peverelli, A., Ruffolo, P., Passarotti, M., Sanna, H., Signoroni, E., Ventura, V., and Zampedri, F. (2019). Nunc Est Aestimandum. Towards an evaluation of the Latin WordNet. In Raffaella Bernardi, et al., editors, *Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, pages 1–8, Bari, Italy. CEUR-WS. org, CEUR-WS.org.

Georges, K. E. and Georges, H. (1913–1918). *Ausführliches lateinisch-deutsches Handwörterbuch*. Hahn, Hannover.

Glare, P. G. (1982). *Oxford Latin Dictionary*. Oxford University Press, Oxford.

Gracia, J., Montiel-Ponsoda, E., Vila-Suero, D., and Aguado-de Cea, G. (2014). Enabling language resources to expose translations as linked data on the web. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 409–413, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Gradenwitz, O. (1904). *Laterculi vocum Latinarum: voces Latinas et a fronte et a tergo ordinandas*. Hirzel, Leipzig.

Khan, A. F. (2018a). Towards the Representation of Etymological Data on the Semantic Web. *Information*, 9(12):304, December.

Khan, F. (2018b). Towards the Representation of Etymological and Diachronic Lexical Data on the Semantic Web. In John P. McCrae, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). event-place: Miyazaki, Japan.

Litta, E., Passarotti, M., and Culy, C. (2016). Formatio formosa est. building a word formation lexicon for latin. In Anna Corazza, et al., editors, *Proceedings of the third italian conference on computational linguistics (clic–it 2016)*, pages 185–189, Naples. aAccademia University Press.

Litta, E., Passarotti, M., and Mambrini, F. (2019). The treatment of word formation in the LiLa knowledge base of linguistic resources for Latin. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 35–43, Prague, Czechia, September. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

Mambrini, F. and Passarotti, M. (2019). Harmonizing different lemmatization strategies for building a knowledge base of linguistic resources for Latin. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 71–80, Florence, Italy, August. Association for Computational Linguistics.

McCrae, J. P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., Osenova, P., Pareja-Lora, A., and Pool, J. (2016). The open linguistics working group: Developing the linguistic linked open data cloud. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2435–2441, Portorož, Slovenia, May. European Language Resources Association (ELRA).

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: development and applications. In *Proceedings of eLex 2017*, pages 587–597.

Minozzi, S. (2010). The Latin WordNet project. In P Anreiter et al., editors, *Latin Linguistics Today. Latin Linguistics Today. Akten des 15. Internationalen Kolloquiums zur Lateinischen Linguistik*, pages 707–716, Innsbruck. Institut für Sprachen und Literaturen der Universität Innsbruck.

Nowak, K. and Bon, B. (2015). medialatinitas.eu. towards shallow integration of lexical, textual and encyclopaedic resources for latin. In *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*, Ljubljana and Brighton. Trojina, Institute for Applied Slovene Studies and Lexical Computing LTD.

Passarotti, M., Budassi, M., Litta, E., and Ruffolo, P. (2017). The Lemlat 3.0 Package for Morphological Analysis of Latin. In Gerlof Bouma et al., editors, *Proceedings of the NoDaLiDa 2017 Workshop on Process-*

ing Historical Language, volume 133, pages 24–31, Gothenburg. Linköping University Electronic Press.

Städtler, T., Dörr, S., Tittel, S., Kiwitt, M., and Möhren, F. (2014). Dictionnaire étymologique de l'ancien français (deaf).

Stead, S., Doerr, M., Ore, C.-E., and Kritsotaki, A. e. a. (2019). Crminf: the argumentation model, version 0.10.1 (draft). `http://new.cidoc-crm.org/crminf/sites/default/files/CRMinf%20ver%2010.1.pdf`. Accessed February 13, 2020.

Tittel, S. and Chiarcos, C. (2018). Historical lexicography of old french and linked open data: transforming the resources of the dictionnaire étymologique de l'ancien français with ontolex-lemon. In *Proceedings of the 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science, Co-Located with LREC2018, Miyazaki, Japan*, volume 12.