# Methodological Aspects of Developing and Managing an Etymological Lexical Resource: Introducing EtymDB 2.0

Clémentine Fourrier, Benoît Sagot

# Methodological Aspects of Developing and Managing an Etymological Lexical Resource: Introducing EtymDB 2.0

**Clémentine Fourrier  Benoît Sagot**
Inria
{clementine.fourrier, benoit.sagot}@inria.fr

**Abstract**

Diachronic lexical information was mostly used in its natural field, historical linguistics, until recently, when promising but not yet conclusive applications to low resource languages machine translation started extending its usage to NLP. There is therefore a new need for fine-grained, large-coverage and accurate etymological lexical resources. In this paper, we propose a set of guidelines to generate such resources, for each step of the life-cycle of an etymological lexicon: creation, update, evaluation, dissemination, and exploitation. To illustrate the guidelines, we introduce EtymDB 2.0, an etymological database automatically generated from the Wiktionary, which contains 1.8 million lexemes, linked by more than 700,000 fine-grained etymological relations, across 2,536 living and dead languages. We also introduce use cases for which EtymDB 2.0 could represent a key resource, such as phylogenetic tree generation, low resource machine translation and medieval languages study.

**Keywords:** Etymological lexicon, Lexical Resource Development, Language Resource Life-cycle, Methodology

## 1. Introduction

Most available electronic lexical resources are synchronic (formalising a language as it is or was at a specific, although sometimes broad, period of time). Until recently, the few diachronic resources available were mostly used for computational historical linguistics tasks (List et al., 2018; Carling et al., 2018). However, over the last few years, diachronic resource use has found its way to more general applications, notably for tasks involving low resource languages, from machine translation using diachronic language relations (Nguyen and Chiang, 2017) or cognates and loan words sets (Grönroos et al., 2018) to bilingual lexicons generation for low resource languages (Nasution et al., 2017).

Yet only a small number of multilingual etymological lexicons are available (de Melo, 2014; Sagot, 2017; Pantaleo et al., 2017; Batsuren et al., 2019), most extracted from the etymological information found in the Wiktionary.[1] There is still room for improvement regarding the quality, richness, lexical or language coverage and etymological granularity (differentiation between inheritance, borrowing, cognacy) of such resources.

The work described in this paper has two parallel objectives. We investigate the methodological challenges underlying the development and use of an etymological database based on the Wiktionary. At the same time, we describe how we addressed these challenges in the case of our own etymological database, EtymDB. After a description of existing diachronic lexicons, we go through the life-cycle of an etymological resource in five steps: creation, update, evaluation, dissemination and exploitation. We use as a case study our previous work on the development of EtymDB's initial version, EtymDB 1.0 (Sagot, 2017), as well as the development of a new version, EtymDB 2.0, its evaluation, dissemination and exploitation. To illustrate the latter, we discuss possible applications in low resource language studies, machine translation, and describe how we used EtymDB 2.0 to extract a global language phylogenetic tree.[2]

Our main contributions are therefore twofold: methodological proposals for the development of etymological lexical resources, and the new EtymDB 2.0 etymological database.[3]

## 2. State of the Art

### 2.1. Existing Etymological Databases

Though a number of individual cognacy datasets can be found online, they often vary wildly in usability, reliability and scope. The following etymological databases combine large scope, usable data format, and generally reliable sources. Interestingly, a majority of these large scale electronic etymological databases have been generated from a version of the Wiktionary, an online collaborative dictionary, which contains large scale structured information, most of the time sourced from already existing and published etymological works.

**EtymWordNet** is an older etymological database extracted from the 2013 version of the Wiktionary (de Melo, 2014). It makes a difference between cognacy[4] and generic "etymological origin" relations, but goes no further. It also does not systematically differentiate between glosses.[5] It contains 473,433 general etymological relations and 538,588 cognacy relations, and was the reference point for EtymDB 1.0 (which was considerably more granular).

[1] http://en.wiktionary.org

[2] A language phylogenetic tree is a speculative rooted acyclic graph displaying evolutionary relationships between languages.

[3] EtymDB 2.0 is available, as its previous version, under a CC-BY-SA free resource licence.

[4] Given two languages with a common ancestor, two words are said to be cognates (in the strictest sense) if they are an evolution of the same word from said ancestor, called their *proto-form*.

[5] Here, the gloss of a word refers its meaning expressed as its English translation

**CogNet** is an automatically extracted cognate database based on wordnets (Batsuren et al., 2019). It uses a loose definition of cognacy, and therefore is actually a database containing both cognates and loanwords. It has the lowest granularity, but the most lexemes, with 3 million "cognate pairs" across 338 languages.

**EtymDB 1.0** is the previous version of our etymological database (Sagot, 2017). It makes a difference between inheritance, borrowing, and cognacy, and contains 1 million distinct lexemes linked by half a million distinct relations. However, it has a few shortcomings, most notably in its management of duplicates. As mentioned in the introduction, both its update and the guidelines we followed for said extension are the topic of the current paper.

**EtyTree** is a graphical etymological dictionary (Pantaleo et al., 2017). It provides information about direct inheritance relations, descendants and compounding, but does not provide cognacy information. Its authors present it more as being a tool to visualise the Wiktionary as it is than a new database in itself, as little extra cleaning and filtering was done on the extracted words. We used this etymological resource as a reference to which we compared EtymDB 2.0, since it was the closest in terms of methods and data source.

For the sake of completeness, it should be noted that **CoBL**, the Cognacy in Basic Lexicon database (Anderson et al., to be published) will probably be a good reference for future etymology databases. It is a descendant and upgrade of Michael Dunn's Indo-European Lexical Cognacy Database, and though it only concerns itself with cognacy, its sources have been handpicked. It is sadly not yet available to the public, and as such we were not able to use it as a reference point for this paper.

### 2.2. Existing Etymological Formats

When talking about open general linguistic data, Chiarcos et al. (2013) highlight the advantages of using a Resource Description Format (RDF), notably in terms of interoperability enhancement. However, for etymological data, until 2019, no single standardised format existed, and the computational historical linguistics community was using one or the other of the following initiatives.

McCrae et al. (2012) introduced the **LExicon Model for ONtologies**, or lemon, which builds on the RDF, for general lexical data. This models separates lexical entry (container for forms and meanings/glosses), lexical form, representation (orthography), lexical sense, and subcomponents. Trying to be as generic as possible and using Semantic Web standards, this model discourages the duplication of information, and encourages instead to share said information across items using a linking mechanism. It was extended in Sérasset (2015) with DBnary, to allow the creation of multilingual resources with the format, and on this, Pantaleo et al. (2017) built another extension, this time specifically designed to manage etymological data. Lemon is a very rich format, with a wide range of applications, and is perfect for interoperability needs; yet this added complexity and superposition of layers prevents it from being a straightforward format for etymological data.

Salmon-Alt (2006) introduced an XML format specifically designed for etymological data, where a basic unit is an etymon (a word located, in time and space, in relation to other words), and relational units are represented as etymological links, between an etymological source and target, typed by etymological classes. Bowers and Romary (2017) extended this model to standardise and integrate it in the **Text Encoding Initiative** (TEI), most notably to include etymological as well as lexical creation processes as relation types: standard inheritance, borrowings, metaphors, metonymy, composition and grammaticalisation.

In 2019, the **Lexical Markup Format** (LMF), the official ISO standard for NLP and digital lexicons (which provides guidelines to model and encode lexical information) has been extended in (Romary et al., 2019) to introduce management of etymological information; the format chosen for its serialisation was the TEI.

## 3. Creating an Etymological Lexicon From Available Datasets

As this section describes, in part, EtymDB 1.0, which has already been introduced by Sagot (2017), only the relevant information for the current article will be summarised here.

### 3.1. Defining Goals

To define goals for a new resource, it is important to know both what is available and the limitations of existing works. When EtymDB 1.0 was first created, the only existing resource among those we described in Section 2.1. was EtymWordNet (de Melo, 2014). As described earlier, it did not differentiate between glosses nor had a fine relation granularity. As such, the initial goal of EtymDB was to provide a large scale formalised etymological database at the lexeme level, which differentiated glosses, and contained a finer level of granularity regarding etymological relations.

### 3.2. Selecting Sources

Getting sourced data is crucial to making a database relevant, especially in etymology: the use of crowdsourcing would make little sense, as it is not possible to just ask bystanders with no etymological expertise to annotate the correctness of etymological relationships. Identifying said relationships is an already highly specialised task, therefore it is very important to know precisely where one's data comes from and how reliable its sources are.

In this light, a number of papers have turned to online collaborative dictionaries such as the Wiktionary, which contain large scale structured information, most of the time sourced from already existing and published etymological works. Meyer and Gurevych (2012) contains a (now slightly dated) full description of the Wiktionary, as well as a discussion of its update mechanism, coverage and quality, in comparison to expert resources. Sérasset (2015) considers it to be an interesting starting point to build linguistic resources from. The Wiktionary has been used in general NLP tasks, such as semantic relatedness assessment (Zesch et al., 2008), cognate clustering (using translation pairs, see Wu and Yarowsky (2018)). It has also been used in linguistic resources creation, such as encyclopedic dictionary and ontology generation (Ehrmann et al., 2014), wordnet induction (de Melo, 2014), or etymological tree representation (Pantaleo et al., 2017). More recently, Hartmann (2019)

argued that "Especially regarding reconstructed language data, Wiktionary has the decisive advantage that the reconstructions follow certain guidelines [...] unlike data collected from various different traditional dictionaries".
This rationale was behind the choice made in Sagot (2017) to use the Wiktionary as the starting point for EtymDB.

## 3.3. Identifying Bias

No matter the source chosen, it will be biased, either due to the data gathering process or to external factors.
In the case of EtymDB, for example, using the Wiktionary biased the language distribution of our dataset towards English. It also biased the language scope: since it is a collaborative dictionary, entries for non English words are either made by specialists or by people speaking several languages including English. As such, languages whose native speakers do not also speak English have few entries in the Wiktionary; this reflects the low resource status of such languages. For example, in October 2019, the Wiktionary contained more than 500,000 entries each for English, Latin, Spanish and Italian; in comparison, more than 3,500 languages had less than 100 entries, among which more than 2,500 have less than 10 entries. Among those, one could find, as expected, dead or proto-languages, critically endangered languages or dialects, such as Baré (2 speakers left in the world, 2 entries in the Wiktionary), as well as languages with a non negligible number of speakers, such as Kabyle (87 entries, 5 million speakers) or Igbo (82 entries, 27 million speakers) that were nonetheless under-represented.

## 3.4. Formalising Lexical Information

If the initial database is of a considerable size and contains extremely varied information, its initial lexical information formalisation is unlikely to be adequate for the chosen goals (as is the case when working with etymological data in the Wiktionary). Therefore, is is important to have identified the initial lexical representation, and defined the target one.

### 3.4.1. Original Wiktionary Lexical Entries

The Wiktionary is structured around lexical pages, which contain all information for a given head word, in the form of one or more lexeme-level lexical entries in one or more languages. Each lexical entry is supposed to contain the lexeme, its language, its part of speech, its definition, a reference, and links to other relevant words. A Wiktionary dump is structured using two formats. The lexical page structure and metadata for each page are encoded in XML. The content of each page uses the wiki format, following templates and typographical markers for representing titles and structured information (such as etymological information). We will have to manage those two different data structures to extract as much relevant information as possible.

### 3.4.2. Defining Etymological Lexical Entries

A multilingual etymological database is composed of lexical entries and of relations between them. The level of granularity needed to represent etymological information must be chosen first: for etymological lexical resources, a lexical entry must be composed at least of a lexical unit, in this case a lexeme (both terms will be used interchangeably), represented as a triplet: a citation form (lemma), glosses (representing meaning), and a language identifier. The resource can also contain extra information about the lexical unit, such as its part of speech. Then, relations between items must be defined: the database can contain both direct or indirect relations, of different types.
For EtymDB 1.0, a lexeme was defined by a citation form in its own language, a language identifier, and English gloss(es). Each lexeme was then associated to a unique numerical identifier (as recommended by Chiarcos et al. (2013)). No extra information was added. The relation between lexemes was then defined as being a directed and direct relation, without intermediaries, between two minimal items. Those relations were classified in 4 categories: inheritance, borrowing, lexical creation (morphological derivation or composition), and cognacy.
To store this data, EtymDB 1.0 introduced its own data structure, separating the above information into three groups: a set of lexical units (lexemes, with their associated languages, glosses when provided, and unique ids), a set of simple etymological relations, defined as one source lexeme and one target lexeme associated with a relation type, and a set of complex etymological relations involving several source lexemes and one target lexeme associated with a relation type (e.g. for compounds).

## 3.5. Extracting Relevant Data

Desquilbet et al. (2019) describe the measures that should be taken in order to ensure reproducibility when working with data, among which recording each and every step of the data processing, ideally in a script.
In the case of EtymDB 1.0, the 2017 Wiktionary data dump was first converted to an homogeneous XML file containing only the relevant information extracted from lexical entries for 831,988 lexemes. The high number of templates and the variations in how they were applied in the initial data contributed to the complexity of the task, which was accomplished using a variety of regular expressions, fitting as many cases as possible. The resulting standardised entries contained a lexeme, the content of the page's 'Etymology' and 'Form' sections, as well as glosses, when available.
This XML file was then parsed in order to extract relation information as well as select individual lexical units. Several challenges were encountered, the main ones being that, first, the etymological data could be present either in an XML tag or in plain English in the text,[6] and second, that a lexeme could be associated to several glosses (a same lexeme can be cited in more than one page with different glosses) or to no glosses at all (a lexeme can be mentioned in the etymological part of a lexical entry without being glossed). To face these challenges, patterns were defined to find new glosses depending on context. Then lexical units, composed of a lexeme, its gloss(es) and its language, as well as etymological relations linking lexemes and associ-

[6]For some languages, such as Chinese, Korean or Japanese, the etymological information was not located in the entries themselves, but within categories assigned to the pages, (such as "Korean terms derived from Middle Korean", which contains 1,049 entries); not all the information in the page categories having been reported to the pages, we expect to have lost etymological information for such languages.

ated with a relation type, were extracted from the XML. Duplicate lexemes caused by glossing variation (lexemes having the exact same orthography and associated to different glosses having common words, or empty glosses) were merged using a number of heuristics. However, duplicates due to formal variation (similar orthography up to the different accentuation or diacritics) were considered different lexemes. Finally, etymological relations were filtered, in order to keep only direct relations between items and indicate as precisely as possible when relations represented borrowings or morphological derivation (for more information see Sagot (2017)). It is important to note that synchronic compounding relations within a language (e.g. English *agroforestry* coming from English *agro-* and English *forestry*) were not kept, as the goal was to create a diachronic database.

This permitted the creation of EtymDB 1.0, a CSV etymological database, divided in two parts: base units, composed of an id, a lexeme, a language id and one or several glosses, and relations, containing the relation type, the target lexeme id and the source lexeme id(s).

## 4. Updating the Resource

Whether it be with new data or by added post-processing and data filtering, it is also possible to create a new resource by updating already existing ones. For EtymDB 2.0, we both update the data and improve its filtering.

### 4.1. Using New Data

In terms of database updating, there are two strategies.

The first one is to update the original database with a newer, bigger version of its data; doing so creates a validation opportunity by enabling the comparison of the new version against its previous form. In the case of EtymDB 2.0, we choose to use a more recent version of the same dataset, the Wiktionary. (The data choice follows the same reasoning as earlier). The 2019/10/20 data dump contains nearly 6.1 million articles, which represented an addition of at least 600,000 articles since the last database version (an increase of almost 10% in size). It is then parsed using the same scripts as the previous version, first creating an intermediate structured information file, then extracting relational data from said file.

The second update strategy is to complete the original database with data from other sources, which can increase its coverage. However, the added data first needs to be mapped into the original database format. It might also intersect with data in the original database, which increases the challenge of dealing with duplicate items or relations. This strategy was used by Chiarcos and Sukhareva (2014), who created their etymological database by linking etymological dictionaries for a number of Germanic languages (converting them into a common RDF format and dealing with duplicates), and then extended it with Wiktionary data.

### 4.2. Improving Data Filtering

Whatever the data origin (various sources, rich databases containing various information), it is likely that duplicate lexemes are present. They can have two causes: glosses variations (due to difference in descriptions), or formal variations (as discussed in Section 3.5.).

For EtymDB, we explained in Section 3.5. a first method to take care of glosses variation by merging different glosses containing the same words. Glosses variation can also be taken care of by using wordnets, as done in (Batsuren et al., 2019).

However, EtymDB 1.0 did not take care of formal variation. For instance, the Greek words παροιμίᾱ and παροιμία, both with gloss 'proverb', were considered as different lexical units because of the difference in vowel length indication on the last letter; similarly, Lithuanian *šaltìnis* 'source' and *šaltinis* 'spring, source, that which is cold' were two lexical units because of the difference in vowel accentuation on the middle "i". This is why, when building EtymDB 2.0, we decide to find such duplicates by merging lexemes whose citation forms only differed by diacritics and whose respective glosses had a non empty intersection (i.e. both lexemes share at least one common gloss). This allows us to merge 415 word pairs, including those cited above.

Both these updating steps, the source data update as well as the improved data filtering, allow us to generate a new version of EtymDB, dubbed EtymDB 2.0.

## 5. Describing and Evaluating the Resource

### 5.1. EtymDB 2.0: Quantitative Information

The initial conversion process from the Wiktionary to our standardised XML file produces more than 2.05 million lexemes, 80,265 lexeme sequences (used in complex etymological relations) and 738,845 etymological relations. After data cleaning, by applying the steps described in Sections 3.5. and 4.2., 170,601 lexeme merging operations are performed, and 4,269 non direct relations are discarded. This results in 1.8 million distinct lexemes linked by 724,906 distinct relations. The relations comprise 155,933 cognation relations and 568,973 relations of another type. The lexemes obtained belong to 2,536 languages, the most represented being English, which constitutes 48% of the database with 911,086 lexemes, Latin (69,224 lexemes), French (34,488), Italian (31,295) and German (27,009). 414 languages are well represented with more than 100 lexemes, whereas 769 languages only have one lexeme. 1,129,032 lexemes (60%) have a gloss.

In Section 7., we will discuss the data more in depth, to illustrate three interesting use cases for etymological resources such as EtymDB. In doing so, we shall show that EtymDB contains useful lexical information on languages which are often poorly resourced otherwise.

### 5.2. Scope: Comparing with a Previous Version

Comparing an updated database with its previous version has several benefits, among which ensuring that the scope actually has been extended, and validating the relevance of the results.

For instance, EtymDB 2.0 contains 50% more lexemes and 40% more relations than EtymDB 1.0 (when the number of articles in the Wiktionary only increased by 10%). Most of these new lexemes are English words (the database contains 3,5 times more English words than before); however, 225 new languages appear in the updated database.

85% of the relations in EtymDB 1.0 are kept in EtymDB 2.0. However, some, change in nature. 75 % of the inheritance relations, 93% of the cognacy relations, and 94% of the borrowing relations are kept as such, but 4% of the initial borrowing relations become inheritance relations, and 3% of the original inheritance relations become borrowings. The disappearing relations have several origins, among which our data filtering. Relations that happen between words from EtymDB 1.0, that were merged in EtymDB 2.0, represented duplicate relations which are removed in the newer version. For example, EtymDB 1.0 contained a cognacy relation between Latvian *vārna* and both Lithuanian *varna* (unaccented) and Lithuanian *várna* (accented); these two relations are merged with our improved filtering, and only the cognacy relation between Latvian *vārna* 'crow' and Lithuanian *várna* remains.

### 5.3. Quality: Validation by Experts

Etymology being a highly specialised field, the best evaluation possible for etymological data is one done by experts, in this case historical linguists and etymologists. Yet, a single linguist or team of linguists will not be able to evaluate an entire etymological database containing several thousands of words within a reasonable time frame. As such, to be analysed, data must first be divided into relevant subsets of a reasonable size. Such subsets can then be compared to prior knowledge from quality sources (etymological dictionaries, reference documents) or to the outcome of the application of established language changes rules. We are not able to apply expert validation to our data but feel that it is important to mention nonetheless.

It is highly unlikely that an automatically extracted database such as EtymDB has a uniform distribution of words and relations across language. Therefore, it is interesting to first provide general statistics to experts (about the overall distribution of languages, the distribution of cognacy/inheritance/borrowing relations for well known languages, and so forth), in order to diagnose some of the biggest problems the base could have. Once the base is established as sound, random samples can be extracted and analysed. Several types of analyses should be considered: diachronic information (etymological relations), synchronic information (words existing synchronically in a given language), relational accuracy on a couple of highly resourced languages (if it is low, it is more likely due to an error in the extraction process than in the data itself), relational accuracy on low-resource languages (which could contain new and interesting information, and highlight the novelty of the database contribution). If problems are identified in the extracted database, the first step is to see whether these issues are present in the source data; if they are, a method to eliminate aberrant data during extraction must be designed. If they are not, the extraction algorithms must be analysed step by step to see where they inject errors.

As etymology is in most cases a field of authority, providing the sources of the compiled etymologies will help linguists to rapidly make a difference between reliable sources which barely need extra checking, and unreliable or absent sources, which will need a more in depth analysis.

### 5.4. Quality: Comparison with Related Work

Even when a manual evaluation can be carried out, automatic comparison to similar existing resources is an important way to assess the quality of a lexical database. As discussed in Section 2.1., the closest resource to EtymDB 2.0 is EtyTree (Pantaleo et al., 2017), as it was produced by extracting etymological and relational information directly from the Wiktionary. The EtyTree data is provided in an extended lemon format, from which we extract its 694,923 words and 889,101 relations.

EtymDB 2.0 contains 22% of EtyTree's words and 8% of its relations. By looking in more detail at the words present in EtyTree but not EtymDB 2.0, we notice that 94% of them are English compounds and English inflected words. This results from a crucial difference between both resources: EtyTree takes into account compounding and inflectional relations within a language, whereas EtymDB 2.0 ignores those and focuses on diachronic information by design. After removing relations from a language to itself from Etytree (compounds, inflections, derivations) from the comparison, EtymDB contains 67% of EtyTree's words and 48% of its relations.

Conversely, EtyTree only contains 24% of EtymDB 2.0's words, and almost none of its relations; again, this can be explained by design differences: EtymDB pays close attention to borrowings and cognates, when EtyTree does not always take into account such information.

The difference in approach between the two bases can be illustrated with an example. In EtymDB 2.0, English *feed* 'to feed' is indicated as inherited from Middle English *feede* 'to feed', cognate with Dutch *voeden* 'to feed', West Frisian *fiede* 'to feed', Danish *føde* 'food', Swedish *föda* 'food' and Icelandic *fæða* 'to give birth to, to feed', as well as borrowed in Portuguese *feed* and Spanish *feed*, both 'Internet feed'. In EtyTree, English *feed* 'to feed' is indicated as inherited from Middle English *feede* 'to feed', having etymological derivatives in English only, among which *feeding frenzy*, *fish feed*, *misfeed*, *underfeeed*, *data feed*, and having etymological children in English only (such as *infeed*, *feedfest*, *multifeed*, *feedyard*).

We conclude that EtyTree is actually more of a morphological derivation and inheritance database, when EtymDB 2.0 is more of an etymological database. EtymDB 2.0's word extraction process seems relevant as it extracts almost as many words as EtyTree. It is harder to judge the quality of the relations as no other base has as fine a granularity as ours, for actual etymological relations.

## 6. Disseminating

To be interesting to the scientific community, a database must be available and usable. As such, it needs to be exportable in a readable format, licensed and documented.

### 6.1. Choosing an Suitable Format

To be usable by the scientific community, a dataset must be provided in a readable and easily usable format. For this reason, it is usually good practice for interoperability and sharing purposes to use a format of reference, which in this case would now be the format from the Text Encoding

Initiative (Romary et al., 2019). In our case, EtymDB is provided as a CSV file, and can also be exported in TEI.

## 6.2. Documenting the Resource

### 6.2.1. General Documentation

Language databases are generally described in a research paper that indicates where the data comes from, how it was extracted and parsed, and how errors were managed. To ensure reproducibility, it is also important to provide the original data, the scripts which were used to generate the data, and a documentation for future users, which details how to use the base, as well as summarises information type, encoding, structure, and ideally sources. All three items are provided along our database on a git repository.[7]

### 6.2.2. Data Statement

Bender and Friedman (2018) proposed the introduction of data statements as "a design solution and professional practice for natural language processing technologists." In concrete terms, the data statement of a dataset is a description of all elements needed to understand the context of its creation and edition, among which the curation rationale, language variety description, and different sub-items depending on the type of data (annotator demographic for annotated data, speaker demographic, speech acquisition situation and recording quality for audio data, text characteristics (such as type and topic) for textual resources...).

In the case of etymological resources, it is important to provide the data provenance, as was done in Section 3., the curation rationale, as in Section 4.2., and the language variety, as we did in Section 5.1.. All this information was synthesised in a Data Statement available with the dataset on git. We also tried to reflect on the data bias, in Section 3.3..

## 6.3. Licensing

Licensing a resource is declaring who can use it, what for, and under which conditions. A resource without an explicit license is legally considered to be under exclusive copyright, and as such not usable by anyone but its authors. Choosing and providing a license with each resource is therefore an crucial step in making the data available to a wider public. Amongst available licences, non-restrictive public open-source distribution licenses allow and encourage the use of resources with very few or no restrictions.

When providing a license for a dataset extracted from existing resources, it is vital to look at their own licensing, as they can impose restrictions on the licenses under which you can distribute your derived database. In our case, the Wiktionary is licensed as CC BY-SA, i.e. anyone can distribute and share EtymDB 2.0 as long as, first, the original resource (the Wiktionary) is mentioned as being the data source, and second, our database is distributed using the same open and non restrictive license, and thus can be used, modified and redistributed, as long as its authors are cited.[8]

[7] https://github.com/clefourrier/EtymDB

[8] The extraction and analysis scripts we developed are also distributed as free software, under the LGPL License.

## 6.4. Providing Visualisation Tools

A database can also be enriched by contributing navigation or visualisation tools, provided as, for example, web pages. It allows the users to look for lexemes/glosses/relation types easily, or display inheritance trees. Two very good examples of this are the Concepticon and EtyTree.

The Concepticon (List et al., 2017) is not an etymological database, but a database linking concepts from the literature varied concept lists. It is provided as a website, which can be queried, and from which data can be downloaded in several formats (CSV, JSON, XML...).

EtyTree (Pantaleo et al., 2017) was introduced earlier, and their website provides an interface to look at a word ancestor from its name, language and gloss. It is only possible to display trees from the bottom up (displaying all the ancestors of a child word) with no fine tuning (when it could be interesting to see all the children of a given word, to display several etymology trees, or to filter ancestors of a word based on language).

EtymDB does not have at the moment a website and representation, but it is planned as a future work, to help the dissemination and validation of our data. However, it is provided with several scripts to facilitate use.

# 7. Exploiting the Resource

The exploitation of etymological lexicons can involve a number of computational linguistic tasks, ranging from automatically detecting cognates to improving machine translation between a low-resourced language and a related, high-resourced one thanks to cognate and/or loan word pairs. Conversely, such use cases can be used to validate the etymological resource itself via task-based evaluation, by looking at how much the use of the resource improves the tasks—a topic that would deserve a paper of its own.

In this section, we introduce three use cases for etymological lexicons such as EtymDB 2.0: phylogenetic tree generation, as well as ancient language studies and low resource language translation for which we only sketch future research directions.

## 7.1. Phylogeny Reconstruction

Linguistic phylogenetic trees constitute a good approximation of the relations between present and past languages, in which a language is related to its immediate attested or reconstructed ancestor, i.e. the language from which it has inherited the most features.[9] By reconstructing a phylogenetic tree from an etymological database, it is possible to evaluate its quality while being an interesting task *per se*. In particular, it can be used to validate and enhance said resource, which we intend to do for our own.

We introduce two algorithms for the automatic extraction of a phylogenetic tree from an etymological database, and apply them to EtymDB 2.0. Both algorithms follow the same

[9] Amongst such features, the lexicon (even restricted to the base lexicon) does not provide the best evidence. Grammatical features are more trustworthy, such as shared irregularities in the morphological system, which often reflect inherited patters. As a result, it is not always the case that most of a language's lexicon is inherited from its ancestor.

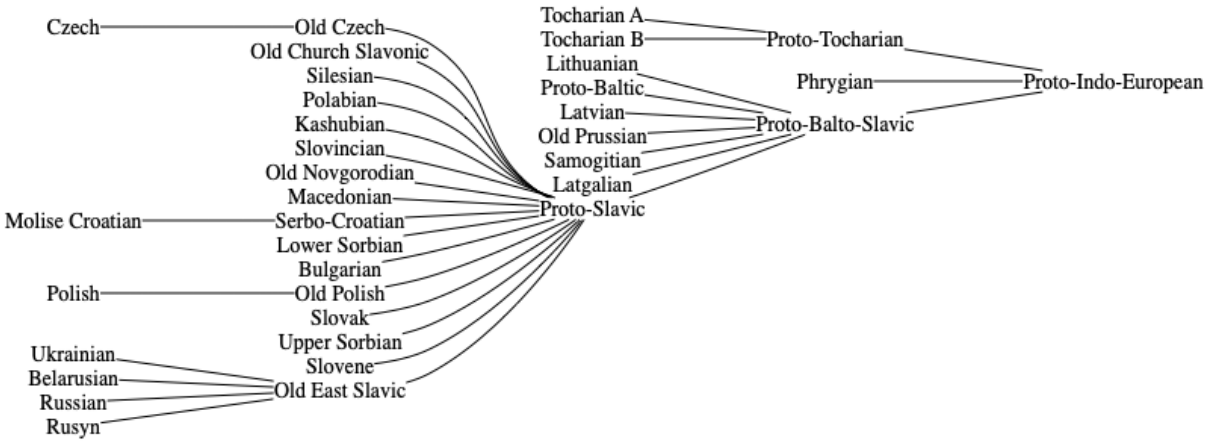


Figure 1: Extract of the tree generated with our refined algorithm

steps: we first create an empty language tree; then, we find all inheritance/borrowing/cognacy relations between two words of two languages of the database, and for each relation, create or reinforce the inheritance link between the two corresponding languages; finally, the resulting set of language relations is filtered, to keep one ancestor per language, thereby producing a set of trees. Our algorithms differ in how this filtering step is carried out.

Our first algorithm, or "naive algorithm", assumes a high quality database, especially regarding inheritance relations. However, since inheritance relations are the default in EtymDB (whereas cognates and borrowings are linked to specific keywords during extraction), we expect this algorithm to encounter several problems. First, borrowings misclassified as inherited words could create erroneous ancestry relations between languages. Second, words are sometimes not (only) linked to their direct ancestors, but (also) to indirect ones. For example, French *étoile* 'star', is linked in our base both with Middle French *estoile* and Latin *stēlla*, when the French term actually descends from Middle French *estoile*, which descends from Old French *estoile*, itself from Latin. We thus developed a second algorithm, or "refined algorithm," to take into account both these problems.[10]

#### 7.1.1. Naive Algorithm

In the naive algorithm, the filtering step is performed as follows: for each language, we keep as its unique ancestor the language connected to it by the highest number of inheritance relations. In order to create a tree with a sufficient coverage, yet small enough to be examined by a human, we discard relations between languages involved in fewer than 20 inheritance relations.

To assess the quality of the resulting phylogenetic tree, we first analyse the overall representation and separation in language families. Our naive algorithm has created several trees rooted in a number of proto-languages and other languages for which no ancestor could be found. Proto-Indo-European, Proto-Austronesian, Proto-Afro-Asiatic, Proto-Turkic, Proto-Uralic are the roots of the biggest trees.[11] Several languages badly covered in our database root smaller trees of their own, often because the relations with their ancestors have been filtered out by our 20-inheritance-relation threshold. It is the case of Ancient Egyptian, for instance (with Demotic as child and Coptic as grand child), whose most frequent ancestor in the database is Proto-Afro-Asiatic, but through only 15 inheritance relations.

When analysing in detail the different trees, both above-described expected problems can be observed. Unsurprisingly, examples of how misclassified relations can affect the phylogeny often involve languages that have heavily borrowed from another one (Latin vs. Greek, Hungarian vs. German). Languages connected to an indirect ancestor often involve situations where the indirect ancestor is better known or studied than the intermediate languages. It is the case for French, for instance, for which EtymDB 2.0 provides more inheritance relations with Latin (2,032 relations) than with its closest ancestors, Middle French (1,311 relations).

#### 7.1.2. Refined Algorithm

In our refined algorithm, we improve the filtering step to try to address these two issues. This step is divided in three substeps.

First, we consider not only inheritance relations, but also cognacy and borrowing relations; we can then remove the link between two languages when they are mostly connected by cognacy or borrowing relations. Therefore, we simplify the tree by first discarding edges where either cognates or borrowings represent more than 25% of the full set of relations.[13] Then, to avoid noise due to wrongly labelled words for low resource data, we remove edges with an inheritance weight lower than five (less than five words in common). We finally keep, for each language, the 5 ancestors connected to it with the biggest inheritance weight. Next, to connect languages to their direct ancestors, we

[10] However, we expect the phylogeny of Chinese, Korean and Japanese to be inaccurate, for the reasons mentioned in footnote 6.

[11] Among the smaller trees ancestors, we notably find Proto-Algic, Proto-Kartvelian, Proto-Thai, Proto-Sino-Tibetan, Proto-Bantu, Proto-Mon-Khmer, Proto-Japonic and Proto-Uto-Aztecan.

[13] The 25% threshold was empirically chosen, after a number of preliminary experiments with various threshold values.

| LANGUAGES | Middle English | Old English | Old French | Middle High German | Middle Low German |
|---|---|---|---|---|---|
| LEXEMES | 20,082 | 14,574 | 13,029 | 5,488 | 3,602 |
| RELATIONS[12] | 22,787 | 20,041 | 18,369 | 8,002 | 3,841 |

Table 1: Number of medieval lexical items contained in EtymDB 2.0

parse the simplified directed graph from bottom to top, choosing for each language its optimal parent as follows. For a given language (e.g. French), we look at its candidate parent list (Middle French - 1,323 relations, Old French - 2,070, Vulgar Latin - 110, etc.). We then look at the possible parents of each candidate parent, which are candidate "grandparents" of the original language. For instance, Middle French has Old French, Latin and Medieval Latin as candidate parents, which are all candidate "grandparents" of French. If a candidate grandparent is also present in the candidate parent list, we remove it from that list. For instance, Old French, now a candidate "grandparent" of French, is removed from the list of French's candidate parents. To take these changes into account in how strong language relations are, we remove for each parent the number of relations from grandparents, and add this count to the child-parent and parent-grandparent relations (as $W_{gp}$). The result is a shorter candidate parent list, hopefully no longer containing grandparent languages.

Thirdly, to pick the best parent in this list, we empirically design a global relation score (to use instead of the simple inheritance count). We want to take into account that the presence of many borrowing relations between two languages indicates strongly that they are not directly related by inheritance (thus we want to penalise it strongly), that the existence of cognacy relations between two languages (indicating inheritance towards common and different ancestors) indicates weakly that one does not descend from the other (thus penalise it weakly), and that the presence of direct inheritance relations and shared grand parents relations should be a good indicator of inheritance relationships.

The final score is the following:

$$R_s = W_{inh} + 2W_{gp} - 20 * W_{bor} - 5 * W_{cog}$$

where $W_{inh/bor/cog}$ are the inheritance/borrowing/cognacy weights of the relation, and $W_{gp}$ is defined above (weights have been empirically adjusted). We then choose the parent with the highest score with regard to the current children.

As expected, a manual analysis reveals that our refined algorithm generates a higher quality tree than our naive algorithm. The links between French, Middle French, Old French, Vulgar Latin and Latin are correctly built this time around. A number of other errors found in the naive tree are absent from the refined tree.[14] However, the graph, though better, is not perfectly accurate, as can be seen in Figure 1 which displays the Phrygian, Proto-Tocharian and Proto-Balto-Slavic branches of said tree; certain parent children relation are close but amiss (Lithuanian, Latvian and Old-Prussian should be descendants of Proto-Baltic, Samogitian of Lithuanian or Proto-Baltic).

[13]Only including inheritance, borrowing and cognacy.

[14]However, languages not well enough or not at all connected to their ancestors in the base, such as Japanese, with only 27 words linked by inheritance to Proto-Japonic (vs. 203 to Proto-Sino-Tibetan) are still misplaced, as expected.

### 7.2. Low Resource Languages Study - Medieval Languages

Medievalists wanting to study such languages usually have to face a scarcity of resources. The Universal Dependency corpus, for example, contains resources for Old French, but none for medieval German or English. EtymDB 2.0 contains 59,000 lexical entries for the combination of Old French, Middle and Old English, as well as Middle High and Middle Low German (against 47,000 for EtymDB 1.0); it also contains 72,000 relations including either one of those languages (for detail, see Table 1, against 57,000 for EtymDB 1.0). As such, EtymDB 2.0 constitutes a valuable historical linguistics lexical resource in itself, as it includes lexical entries for past words, as well as their relations to previous and future lexemes.

### 7.3. Low Resource Languages Translation - Indo-Aryan Family

In the last couple of years, machine translation (MT) researchers have tried to improve the MT of low-resource language pairs using a number of techniques. Bawden et al. (2019) tried to improve MT from English to Gujarati by using Hindi as a pivot language. Grönroos et al. (2018) have used cognates and borrowings to improve MT systems for low-resource languages, by bootstrapping them with the help of etymologically related high resource language translations. EtymDB 2.0 contains 2,316 lexical entries in Gujarati, 7,748 in Hindi and 225 direct cognacy relations between those two languages. It also contains 10,826 lexemes in Sanskrit, as well as 500 etymological relations between Gujarati and Sanskrit and about 3,000 between Hindi and Sanskrit, which could permit the discovery of new cognates between those two languages. Such relations could play an important role in MT for low resource languages.

## 8. Conclusion

We introduced EtymDB 2.0, an etymological lexical resource automatically extracted from the Wiktionary. This resource contains about 1.8 million lexemes in 2536 living and ancient languages, linked by 700,000 fine-grained etymological relations. Over 400 of the languages covered contain information about more than 100 unique lexemes.

Its whole generation, update and validation processes also allowed us to formalise good practices for the development of etymological resources. Beyond development, we also described important resource management aspects, especially regarding dissemination (documenting, providing a data statement, choosing a format and a license).

## Acknowledgements

This work was partly funded by the second author's chair in the PRAIRIE institute, funded by the French national ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001.

## 9. Bibliographical References

Batsuren, K., Bella, G., and Giunchiglia, F. (2019). CogNet: A large-scale cognate database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145, Florence, Italy. Association for Computational Linguistics.

Bawden, R., Bogoychev, N., Germann, U., Grundkiewicz, R., Kirefu, F., Miceli Barone, A. V., and Birch, A. (2019). The university of Edinburgh's submissions to the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 103–115, Florence, Italy, August. Association for Computational Linguistics.

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bowers, J. and Romary, L. (2017). Deep encoding of etymological information in TEI. *Journal of the Text Encoding Initiative*, (10).

Carling, G., Larsson, F., Cathcart, C. A., Johansson, N., Holmer, A., Round, E., and Verhoeven, R. (2018). Diachronic atlas of comparative linguistics (diacl)—a database for ancient language typology. *PLOS ONE*, 13(10):1–20.

Chiarcos, C. and Sukhareva, M. (2014). Linking etymological databases. a case study in germanic. In *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, page 41.

Chiarcos, C., McCrae, J. P., Cimiano, P., and Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*.

de Melo, G. (2014). Etymological Wordnet: Tracing the history of words. In Nicoletta Calzolari, et al., editors, *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, pages 1148–1154, Paris, France. European Language Resources Association (ELRA).

Desquilbet, L., Granger, S., Hejblum, B., Legrand, A., Pernot, P., Rougier, N. P., de Castro Guerra, E., Courbin-Coulaud, M., Duvaux, L., Gravier, P., Le Campion, G., Roux, S., and Santos, F. (2019). *Towards reproducible research*. Unité régionale de formation à l'information scientifique et technique de Bordeaux.

Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J. P., Cimiano, P., and Navigli, R. (2014). Representing multilingual data as linked data: the case of BabelNet 2.0. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 401–408, Reykjavik, Iceland. European Language Resources Association (ELRA).

Grönroos, S.-A., Virpioja, S., and Kurimo, M. (2018). Cognate-aware morphological segmentation for multilingual neural translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 386–393, Belgium, Brussels. Association for Computational Linguistics.

Hartmann, F. (2019). Predicting historical phonetic features using deep neural networks: A case study of the phonetic system of proto-Indo-European. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 98–108, Florence, Italy. Association for Computational Linguistics.

List, J.-M., Walworth, M., Greenhill, S. J., Tresoldi, T., and Forkel, R. (2018). Sequence comparison in computational historical linguistics. *Journal of Language Evolution*, 3(2):130–144.

McCrae, J., Aguado-de Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. (2012). Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46(4):701–719.

Meyer, C. and Gurevych, I. (2012). Wiktionary: A new rival for expert-built lexicons? exploring the possibilities of collaborative lexicography.

Nasution, A. H., Murakami, Y., and Ishida, T. (2017). A generalized constraint approach to bilingual dictionary induction for low-resource language families. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(2):9:1–9:29.

Nguyen, T. Q. and Chiang, D. (2017). Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Pantaleo, E., Anelli, V. W., Di Noia, T., and Sérasset, G. (2017). Etytree: A graphical and interactive etymology dictionary based on wiktionary. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 1635–1640, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Romary, L., Khemakhem, M., Khan, F., Bowers, J., Calzolari, N., George, M., Pet, M., and Bański, P. (2019). LMF Reloaded. In *AsiaLex 2019: Past, Present and Future*, Istanbul, Turkey.

Sagot, B. (2017). Extracting an Etymological Database from Wiktionary. In *Electronic Lexicography in the 21st century (eLex 2017)*, pages 716–728, Leiden, Netherlands.

Salmon-Alt, S. (2006). Data structures for etymology: towards an etymological lexical network. *Bulletin de linguistique appliquée et générale*, 31:1–12.

Sérasset, G. (2015). DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF . volume 6 of *Multilingual Linked Open Data*, pages 355–361. IOS Press.

Wu, W. and Yarowsky, D. (2018). Creating Large-

Scale Multilingual Cognate Tables. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Zesch, T., Müller, C., and Gurevych, I. (2008). Using wiktionary for computing semantic relatedness. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, pages 861–866. AAAI Press.

## 10. Language Resource References

Anderson, Cormac and Heggarty, Paul and The CoBL Consortium. (to be published). *Cognacy in Basic Lexicon database*.

Batsuren, Khuyagbaatar and Bella, Gabor and Giunchiglia, Fausto. (2019). *CogNet: A Large-Scale Cognate Database*.

Gerard de Melo. (2014). *EtymWordNet*.

Johann Mattis List and Christoph Rzymski and Simon Greenhill and Nathanael Schweikhard and Kristina Pianykh and Robert Forkel. (2017). *Concepticon*. ISLRN 045-571-692-786-3.

Pantaleo, Ester and Anelli, Vito Walter and Di Noia, Tommaso and Sérasset, Gilles. (2017). *EtyTree*.

Sagot, Benoît. (2017). *EtymDB*. 1.0.