Extracting an Etymological Database from Wiktionary

Benoît Sagot

Extracting an Etymological Database from Wiktionary

Benoît Sagot

Inria
2 rue Simone Iff, 75012 Paris, France
E-mail: benoit.sagot@inria.fr

Abstract

Electronic lexical resources almost never contain etymological information. The availability of such information, if properly formalised, could open up the possibility of developing automatic tools targeted towards historical and comparative linguistics, as well as significantly improving the automatic processing of ancient languages. We describe here the process we implemented for extracting etymological data from the etymological notices found in Wiktionary. We have produced a multilingual database of nearly one million lexemes and a database of more than half a million etymological relations between lexemes.

Keywords: Lexical resource development; etymology; Wiktionary

1. Introduction

Electronic lexical resources used in the fields of natural language processing and computational linguistics are almost exclusively synchronic resources; they mostly include information about inflectional, derivational, syntactic, semantic or even pragmatic properties of their entries. Because this information is formalised, it can be used by automatic tools.

Conversely, diachronic information such as etymology is virtually absent from electronic resources, only being present in printed or online dictionaries. The few exceptions, such as *The Tower of Babel* database[1] or the *PIElexicon* project,[2] often rely on comparative and etymological principles that are, at best, obsolete or non-consensual,[3] and, at worst, unanimously rejected by the scientific community.[4] Only EtymWordNet (de Melo, 2014), to which we shall come back below, is an outlier in this regard, although it has other severe limitations.

The availability of formalised, detailed and large-coverage etymological databases would make it possible to develop automatic tools targeted towards historical and comparative linguistics. Modelling language evolution and reconstructing proto-languages—ancestors of attested languages—rely on a very large amount of lexical information often covering dozens, if not hundreds of languages. For some language families, such as Indo-European or Semitic languages, almost two centuries of careful work has resulted in a fairly clear understanding of lexical diachrony. However, even for these two families, and *a fortiori* for all others, many grey areas remain.

The development of automatic means to explore possible formal and semantic correspondences between words from different languages and to model their diachronic evolution

[1] http://starling.rinet.ru/babel.php?lan=en

[2] http://pielexicon.hum.helsinki.fi

[3] The Indo-European database in *The Tower of Babel* is based on the Pokorny dictionary, which is now outdated. Moreover, the authors of this database defend non-consensual views on genetic relationships between traditional language families. These views are generally rejected by the scientific community yet still influence some of their etymological proposals.

[4] This applies to the *PIElexicon*, although the justification of such a statement lies beyond the scope of this paper.

would therefore constitute an important step forward for the linguistic sub-fields involved, while raising difficult algorithmic challenges. It would also contribute to the development of resources and tools for the automatic processing of documents written in older forms of the languages, for which they already exist for their modern variant (for instance, documents in Old or Middle English, which cannot be properly processed by tools dedicated to contemporary English). This direction of research should take advantage of the outcome of previous etymological investigations, which should therefore be encoded in the form of formalised electronic lexical resources.

To achieve this goal, we need to find a large-scale source of etymological information, to automatically extract this information from it, and to represent it in a structured or even normalised form. In this paper, we describe a first attempt at carrying out such an enterprise. We rely on the (English) *Wiktionary*,[5] an online collaborative dictionary, whose syntax is semi-structured and which includes relatively detailed and fairly reliable etymological information.

The remainder of this paper is structured as follows. After a brief overview of previous work related to ours (Section 2) and a brief sketch of the various types of etymological relations between lexemes (Section 3), we describe how etymological information is represented in Wiktionary articles and how we extracted and partially structured this information (Section 4). In Section 5, we explain how we transformed this information into a database of lexemes and a database of etymological relations between these lexemes. We provide in Section 6 quantitative information about these two databases, their export formats—including an etymology-oriented extension of the LMF standard[6] currently under discussion—and a manual evaluation of their quality. We conclude this work in Section 7 by discussing possible follow-ups to this work, including possible direct use cases for our etymological databases.

Both databases are freely available under an LGPL-LR license.

2. Related work

Previous work related to ours is threefold: efforts towards the standardisation of etymological information, development of existing databases, and the above-mentioned EtymWordNet.

Since etymological information is only exceptionally taken into account in electronic lexical resources, their structured representation is not yet the subject of recommendations concerning their standardisation. In this regard, the working paper published by Bowers & Romary (2016) reflects the state of ongoing research. It builds on several previous initiatives, including the work by Salmon-Alt (2006). It proposes a set of general principles for the representation of etymological information in electronic dictionaries encoded in TEI. It is based on a relatively broad typology of the underlying phenomena, which covers standard inheritance (what etymologists refer to as *recto itinere*, "in direct line"), borrowing, metaphor, metonymy, composition and grammaticalisation. Some of these mechanisms are not etymological in nature, but are rather lexical creation mechanisms. We shall come back in Section 6 on several limitations of this proposal in its current state.

[5] https://en.Wiktionary.org/
[6] Lexical Markup Framework. See below for details.

Few freely available electronic dictionaries make use of structured representations of etymological information. We have already mentioned *The Tower of Babel* and the *PIElexicon*. Another example is the *Germanic Lexicon Project*[7] by S. Crist, whose representation format can also be considered as a predecessor of the propositions made by Bowers & Romary (2016). However, the various free dictionaries distributed in this framework are only weakly structured: the systematic extraction of etymological relations would be a non-trivial task. This is not the case for the *World Loanword Database*, which, for 1,460 carefully selected meanings, provides one or more lexemes in 41 languages, each associated with a probability level that it results from a borrowing, as well as its possible source lexeme. But the inventory of the 41 languages covered reflects the typological and non-etymological positioning of the project. In any case, it is far from a widely covered resource, and, of course, only borrowing mechanisms are covered, to the exclusion of any other etymological mechanism.

Closer to our work, de Melo (2014) has made available EtymWordNet, which, like in our work, was automatically extracted from the Wiktionary (although in a three-year-old version). However, and despite extensive coverage, the EtymWordNet can not be used as it is for the computerisation of comparative and historical linguistics because of two fundamental limitations. Firstly, the mechanisms at play are not distinguished (for example, no distinction between inheritance, borrowing and morphological derivation). Secondly, and even more importantly, its basic units are lemmas, not lexemes: senses are ignored.

We are not aware of previous works that resulted in a large-scale formalised etymological database at the lexeme level, as is necessary in etymological lexicology (see Section 3) and makes a distinction between etymological mechanisms. That is the purpose of our work.

3. Etymological and lexical creation mechanisms

The extraction and formalisation of etymological information requires a model of this type of information. The first question that arises is that of the basic unit. As recalled by Buchi (2016: 346), only the lexeme can play this role—in our case a lexeme is defined by a citation form, a language identifier and an English gloss.[8] A relation between a lexeme and another lexeme can correspond to changes in the language (diachronic change in the case of inheritance, synchronic change in the case of a borrowing), the citation form (phonetic but also morphological changes) and the meaning (semantic shifts).

The second important question is the nature of the etymological relations between lexemes. Following again Buchi (2016: 346–347), an elementary etymological relation must concern directly related lexemes: there should not be any intermediate lexemes between those involved in the relation. In the case of a *recto itinere* inheritance, and given an inventory of language identifiers, an elementary relation must therefore involve a lexeme in a given language and another lexeme, or several lexemes, in the immediately preceding language or language state.[9] In the case of a borrowing, a direct relation simply involves

[7] http://lexicon.ff.cuni.cz/texts/pgmc_torp_about.html

[8] This also covers the case of place names, person names, people/tribe names, and other proper names.

[9] Fr. *manger* < Mid. Fr. *manger* is therefore a direct relation, contrarily to Fr. *manger* < Old Fr. *mengier* and Fr. *manger* < Late Lat. *manducāre*, which are indirect relations.

the target lexeme and its source lexeme.[10] Using direct relations whenever possible is necessary to be able to correctly specify the nature of the etymological relation involved.[11]

The third question to address for formalising etymological relations is that of the different types of etymological mechanisms. Although we do not cover all of them in this work, we make use of the following typology:

- Inheritance (with phonetic change in most cases, with or without semantic or morphological change); as is customary, we shall note this type of relation as follows: *target lexeme* < *source lexeme*;
- Borrowing; we shall note this type of relation as follows: *target lexeme* ← *source lexeme*;[12]
- Lexical creation
  - Morphological derivation
    - Suffixal derivation; it will be noted as follows: *target lexeme* $<_s$ *base* + *suffix*;
    - Prefixal derivation; it will be noted as follows: *target lexeme* $<_p$ *prefix* + *base*;
    - Other cases (including analogy-based derivation); they will be noted as follows: *target lexeme* $<_d$ *element* + ... + *element*;
  - Morphological composition, noted as follows: *target lexeme* $<_c$ *component* + ... + *component*;
  - Portmanteau word creation, not covered in this work;
  - Truncation and other phenomena, not covered in this work.

To this inventory we shall add a special cognation relation, which will allow us to relate two lexemes (within the same language or in two different languages) that have a common or partly common etymology (in general, at least a same "root"). It will be noted $lexeme_1$ // $lexeme_2$.

4. Extraction and structuration of Wiktionary's etymological information

4.1 Etymological information in Wiktionary

Wiktionary is a collaborative multilingual dictionary. It is organised into articles, which each contain one or more homonymous lexical entries[13] concerning lexemes from one or more languages.

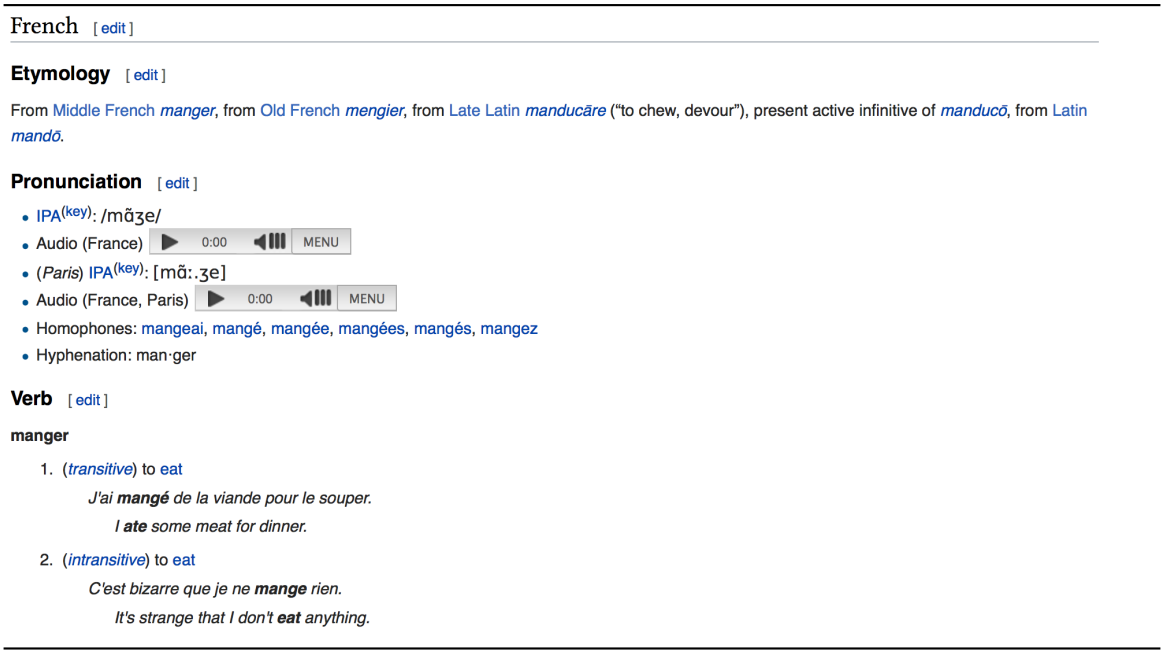
We used the 01/01/2017 dump. It contains nearly 5.5 million articles, more than 40,000 of which are redirect pages. These entries contain a total of 894,453 etymological records.

[10] For instance, relations such as Fr. *abricot* 'apricot' < Esp. *albaricoque* 'id.' and Fr. *abricot* < Port. *albricoque* 'id.' are possible direct relations (both are plausible). The Spanish and Portuguese words are borrowings from Ar. *al-barqūq* 'id.', itself a borrowing from Med. Gr. βερικόκκια 'apricot tree', derived from Ancient Gr. πραικόκιον 'apricot', itself a borrowing from Lat. *praecoquum* 'early (fruit)'. Therefore, a relation such as Fr. *abricot* < Lat. *praecoquum* would be correct but not direct.

[11] Going back to the example introduced in the previous footnote, it would be difficult to assign a simple type to the etymological relation between Lat. *praecoquum* et Fr. *abricot*, as it covers several steps of different natures.

[12] We include in this category all cases of learned loans such as Fr. *oculaire* 'ocular' ← Lat. *ocularis* 'id.'.

[13] Two lexical entries are homonymous if they share the same citation form, independently of the language or part-of-speech of the two underlying lexemes.



French [edit]

Etymology [edit]

From Middle French *manger*, from Old French *mengier*, from Late Latin *manducāre* ("to chew, devour"), present active infinitive of *manducō*, from Latin *mandō*.

Pronunciation [edit]

- IPA(key): /mɑ̃ʒe/
- Audio (France) 0:00 MENU
- (*Paris*) IPA(key): [mɑ̃ː.ʒe]
- Audio (France, Paris) 0:00 MENU
- Homophones: mangeai, mangé, mangée, mangées, mangés, mangez
- Hyphenation: man·ger

Verb [edit]

manger

1. (*transitive*) to eat
   *J'ai **mangé** de la viande pour le souper.*
   *I **ate** some meat for dinner.*
2. (*intransitive*) to eat
   *C'est bizarre que je ne **mange** rien.*
   *It's strange that I don't **eat** anything.*

Figure 1: Part of a Wiktionary entry

This dump is in a semi-structured format: the structuration into articles is encoded in XML and includes metadata for each article; the content of each article is coded using the so-called "wiki syntax", in which the plain text is supplemented by typographical markers (different levels of titles, lists, etc.) and templates allowing the coding of certain information in a systematic way. For example, the template `link` (or `l`) can be used to encode a form that is a link to the article it is the title of. Thus, `{{link|fr|chaise||chair|g=f}}` will be rendered on the Wiktionary site as chaise *f* ("chair"), where the feminine gender is indicated (`g=f`) and where the word chair is a hyperlink to the section of the Wiktionary article "chaise" concerning the French lexems (`fr`).[14]

Figure 1 shows part of the Wiktionary article "*manger*". The corresponding source code is shown in Figure 2.

Finally, "Descendants" sections are sometimes included. They list the descendants of the lexeme at hand, without any further precision on the nature of the etymological relation (inheritance, borrowing).

[14] The language inventory used by Wiktionary is based on the ISO-639-1 to ISO-639-3 standards, with extensions when needed. For more details, cf. https://en.Wiktionary.org/wiki/Wiktionary:Languages. Based on the correspondence between language codes and language names, we also set up a system for the automatic abbreviation of language names as well as a system for the identification of language (codes) based on their usual names or abbreviations as used in the Wiktionary articles. Thus, "OFr.", "Old Fr." or "Old French" are correctly interpreted as reflecting the `fro` language code, which can then be transformed into its standard English abbreviation, " OFr. "

```
==French==

===Etymology===
From {{inh|fr|frm|manger}}, from {{inh|fr|fro|mengier}}, from {{inh|fr|LL.|manducāre||to chew,
devour}}, present active infinitive of {{m|la|manducō}}, from {{inh|fr|la|mandō}}.

(...)

===Verb===
{{fr-verb}}

# {{lb|fr|transitive}} to [[eat]]
#: ''J'ai '''mangé''' de la viande pour le souper.''
#:: ''I '''ate''' some meat for dinner.''
# {{lb|fr|intransitive}} to [[eat]]
#: ''C'est bizarre que je ne '''mange''' rien.''
#:: ''It's strange that I don't '''eat''' anything.''
#: '''''Manger''' au restaurant.''
#:: ''To '''eat''' in a restaurant.''
```

Figure 2: Source code corresponding to the article part shown in Figure 1 (the pronunciation-related part is not shown)

4.2 Extraction and structuration

We first converted the Wiktionary dump into an XML file using a series of regular expressions.[15] This XML file is a set of lexical entries that correspond approximately to lexemes. It contains only entries for which Wiktionary provides etymological information in a dedicated section. It contains is 831,988 entries. Each of them includes the content of this etymological section in an `<etymology/>` tag, in which all forms, especially but not only those mentioned using *templates*, are represented by an XML element `<form/>`. Whenever several `<form/>` are used together (affixed derivation, composition), their combination is harmonised using the symbol "+" (see above). Whenever several alternate forms are listed (variants, principal parts...), they are separated using the symbol "∼". These apparently simple standardisation steps are made complex by the variety of situations, the richness of the available templates and the multiplicity of ways used by Wiktionary contributors to represent etymological information.

If a section listing descendants is available, they are all converted into `<form/>` elements and are included in a dedicated `<descendants/>` element within the `<etymology/>`.

All forms mentioned in the article but outside the etymological section or the descendant section are also extracted in a special section `<forms/>`. This is because these forms, especially those associated with a gloss, might prove useful in the next steps of the extraction process.

Whenever possible, the lexeme at hand is associated with a gloss. If it is an English lexeme, its citation form is considered as its own gloss. In all other cases, we try to extract one or several glosses based on the definitions provided in the article.

[15] This XML format is a working format. It is not intended at this stage to be suitable for TEI compatibility. We shall return in Section 6.2 on how we exported etymological information to an extended TEI format.

From the source code corresponding to the French verb *manger*, shown in Figure 2, our structuration process outputs the entry given in Figure 3.

```
<entry id="manger#French">
  <header><form lang="Fr." l="fr" sense="to eat; food; foodstuff">manger</form></header>
  <etymology>
From <form lang="MFr." l="frm" trgl="fr" trglang="Fr." type="inherited">manger</form>, from <form
lang="OFr." l="fro" trgl="fr" trglang="Fr." type="inherited">mengier</form>, from <form lang="LL."
l="la-lat" sense="to chew, devour" trgl="fr" trglang="Fr." type="inherited">manducāre</form>, present
active infinitive of <form lang="Lat." l="la">manducō</form>, from <form lang="Lat." l="la" trgl="fr"
trglang="Fr." type="inherited">mandō</form>.
  </etymology>
  <forms>
    <form lang="Fr." l="fr">gramen</form>
    <form lang="Fr." l="fr">magner</form>
  </forms>
</entry>
```

Figure 3: Output of our structuration process for the input source code shown in Figure 2

5. Construction of the etymological database

The output of the structuration process described in the previous section is much easier to exploit than the original Wiktionary dump. However, several challenges remain. The main one is of course that the etymological information is given in plain English, apart from the `<form/>` elements. Another one is that, from one article to the other, a same lexeme can associated with different glosses, if any.

To address these challenges, we proceed in several steps. First, we defined a number of regular patterns for infering the gloss of a non-glossed form based on its context.[16] In such a case, the corresponding `<form/>` element is updated accordingly.

We then process all entries and the etymological information they contain, in order to create triples of the form (target lexeme, source lexeme or source lexeme sequence, type of the relation). We now have to merge synonymous lexemes as much as possible. For instance, if the triples we built involve lexemes such as Fr. *bêtement* 'stupidly, idiotically', Fr. *bêtement* '(no gloss)' and Fr. *bêtement* 'stupidly, foolishly', these three lexemes need to be merged into a lexeme Fr. *bêtement* 'stupidly, idiotically, foolishy'. To achieve this goal, we iterate the following steps until stability:

- If a gloss-less lexeme has the same language and the same citation form as exactly one (glossed) lexeme, then these lexemes are merged.
- If two glossed lexemes have the same language, the same citation form, and at least one gloss in common (cf. 'stupidly, foolishly' vs. 'stupidly, idiotically'), then they are merged (in this example, it creates a lexeme with the gloss 'stupidly, foolishly, idiotically', as mentioned above);
- All triples encoding etymological relations are then updated accordingly.

[16] Coming back to our French running example *manger*, the phrase "From Middle French *manger*, from Old French *mengier*...", although it contains no glosses, makes it possible to associate the gloss of the head lexeme *manger*, namely 'to eat', to MFr. *manger* et OFr. *mangier*.

In order to restrict as much as possible our set of etymological relations to direct relations, we remove any relation between two lexemes $lexeme_1$ et $lexeme_3$ such that there exists a relation between $lexeme_1$ and an intermediate lexeme $lexeme_2$ and a relation between this $lexeme_2$ and $lexeme_3$.[17]

Finaly, the type of certain relations is corrected, in order to indicate as precisely as possible cases of borrowing or morphological derivation rather than inheritance, this latter case still remaining the default one.

The outcome of this extraction process is twofold: a set of lexemes, only some of them being glossed, and a set of etymological relations involving a target lexeme, one or more source lexemes (two or more in case of composition or affixal derivations) and a relation type. Here are a few real examples concerning French lexemes:

- Fr. *gobelet* 'goblet' < OFr. *gobel* 'goblet; cup; beaker; tumbler'
- Fr. *maudire* 'to curse' < OFr. *maudire* ∼ *maldire* 'to curse'
- Fr. *éponger* 'to sponge; to absorb' $<_s$ Fr. *éponge* 'sponge' + Fr. *-er*
- Fr. *idéologie* 'ideology' $<_d$ Fr. *idéo-* + Fr. *-logie*
- Fr. *acajou* 'cashew' ← Port. *acajú* 'cashew'
- Fr. *car* 'car; coach' ← E *car*

6. Results and evaluation

6.1 Quantitative information

The initial extraction process described at the beginning of Section 5 has produced almost 1.2 million lexemes, 62,056 lexeme sequences and 548,935 etymological relations between two lexemes or between a lexeme and a sequence of lexemes.[18] A few dozen iterations of the lexeme merging algorithm merged 199,185 lexemes and 289 lexeme sequences, resulting in 975,473 distinct lexemes, 61,809 distinct lexeme sequences and 519,348 distinct relations. After discarding 5,149 non-diret relations, the final number of relations is 514,199.

The lexemes obtained belong to 2311 distinct languages, the best represented of which are, in decreasing order, English (257,978 lexemes), Latin (65,981), French (32,044), Italian (28,028), and Ancient Greek (21,077). Among these lexemes, 659,567 (68%) have a gloss.

Among the 514,199 relations, 452,041 relate two lexemes, whereas other relations relate a lexeme and a lexeme sequence. There are 90,511, cognation relations, all other 423,673 relations being direct relations. Finally, 318,883 relations involved glossed lexemes only.

Note that we could have easily created many more cognation relations by adding relations sharing a (direct or indirect) ancestor in our database.

6.2 Etymological chain inference and TEI export

We have developed an export module for our etymological relation database that encodes data in the TEI format proposed by Bowers & Romary (2016). In this format, direct

[17] The same mechanism applies when $lexeme_3$ is not a unique lexeme but a sequence of lexemes.
[18] In these figures and in all figures below, lexemes invovled in zero relations are not counted.

relations wan be exported in the form of simple <etym/> elements, associated with the type of the relation at hand.

For a given lexeme, it can also be interesting to have not only its direct etymon but also its etymological history in as exhaustive a way as possible. In fact, such an etymological chain is often provided in the etymological information included in Wiktionary articles, as exemplified in Figures 1 and 2. In order to re-build these etymological chains (or derivations) from our relation database, one can simply recursively retrieve relations involving each etymon involved. For instance, from Fr. *manger* 'to eat' < MFr. *manger* 'to eat' and MFr. *manger* 'to eat' < OFr. *mengier* 'to eat', one can re-build the chain Fr. *manger* 'to eat' < MFr. *manger* 'to eat' < OFr. *mengier*. This is how we created etymological chains, before encoding them in TEI.

We had to extend the proposal by Bowers & Romary (2016) in four directions, which could serve as a source of inspiration for its further improvement:

- This proposition does not cover the cogation relation. We therefore introduce an additional relation type (*type="cognate"*) to the element <etym/>.
- It does not allow to refer to another lexical entry providing relevant etymological relation. We therefore introduce a new type (*type="reference"*) to the <etym/> element, within which a direct reference to the relevant lexical entry can be included with an <xr/> (TEI element used for cross-references).
- Bowers & Romary (2016) do not provide any way to encode etymological chains. We simply used a special <etym/> element, which, using a dedicated attribute, indicates that it contains a sequence of etymological relations, each of them being represented by a specific <etym/> element within the global <etym/> element.
- In their document, Bowers & Romary (2016) do not allow for alternative etymological hypotheses, something which is frequent in our database. In this case, we also make use of a special <etym/> element, which indicates using a dedicated attribute that it contains alternative hypotheses, each of them represented by a distinct <etym/>.

In the two last cases, the recursivity of <etym/> elements allows for any possible combinations, such as an etymological chain starting with two "certain" steps followed by an alternative between two different etymological sub-chains.

6.3 Manual evaluation

The evaluation of our etymological relation database could be carried out with four different questions in mind:

1. What is the quality of the etymological information provided by the Wiktionary?
2. What are the errors caused by our extraction and structuration process?
3. What are the errors introduced by our gloss inference and lexeme merging algorithms? Conversely, what is the coverage of these algorithms?
4. Which errors result from the fact that non-typed relations are interpreted by default as inheritance relations?

A detailed answer to the first question is not straightforward, and falls beyond the scope of this paper. An informal study of the etymological information found in a random set

```
<entry xml:id="sla-pro:gostinŭ:guest" xml:lang="sla-pro">
  <form type="lemma">
    <orth>gostinŭ</orth>
  </form>
  <sense>
    <cit type="translation" xml:lang="en">
      <oRef>guest<oRef>
    </cit>
  </sense>
  <etym type="suffixalDerivation">
    <cit type="etymon">
      <oRef xml:lang="sla-pro">gostĭ</oRef>
      <gloss>guest</gloss>
      <etym type="inheritance">
        <cit type="etymon">
          <oRef xml:lang="ine-pro">gʰóstis</oRef>
          <gloss>stranger, guest, host, someone with whom one has reciprocal duties of
                 hospitality</gloss>
        </cit>
      </etym>
    </cit>
    <cit type="etymon">
      <oRef xml:lang="sla-pro">-inŭ</oRef>
    </cit>
  </etym>
</entry>
```

Figure 4: Example of a TEI-formatted entry (cognation relations are not shown)

of articles showed that this information is usually reliable. Only Proto-Indo-European etyma sometimes reflect of a somewhat obsolete knowledge of the field. Nevertheless, it can be considered that etymological information in Wiktionary can generally be trusted, and often reflect the most recent and consensual scientific literature, which are often cited in the references.

The precision and recall of our gloss inference and lexeme merging algorithms are easier to evaluate. We first focused on the recall of the merging algorithm. We randomly selected 50 (language, citation form) pairs among the 124,775 ones (out of 941,757) that correspond to more than one entries. We then extracted all entries for these 50 pairs, and have manually annotated the relevance of their co-existence (as opposed to merging them). In almost all cases, additional merges would have been relevant, but our algorithm was not able to perform these merges. It is therefore an obvious direction for future improvements. Conversely, in order to evaluate the precision of our merging algorithm and that of our gloss inference algorithm, we randomly extracted 100 glossed forms and checked the quality and coherence of their glosses. Out of these 100, we identified two extraction errors (both caused by an unusual use of the "wiki" syntax by contributors), a partial error (some of the glosses are correct, one of them is an easily dismissable "wiki" code fragment), a transcription misinsterpreted as a gloss, and a (correct) definition misinterpreted as a gloss. All other glossed forms were fully correct. Therefore, there are only a few errors, which are almost never caused by our merging and glossing algorithms—yet the extraction and structuration algorithm could be slightly improved.

Finally, we evaluated the etymological relations themselves based on a random set of size 100. Among them, 78 are correct, 18 have type "inheritance" whereas they encode

borrowings, three of them have other relation typing errors, and only one is errineous because of an error while extracting the article. The 18 errors of type "inheritance" instead of "borrowing" are the result of the fact that inheritance is the default relation type, used when the latter is not explicitely provided in the Wiktionary. A finer description of the relations between languages would make it possible to automatically correct these examples. This is something we will do in the near future.

6.4 Comparison with EtymWordnet

EtymWordNet de Melo (2014), freely available without an explicit license,[19] is an etymological database extracted from Wiktionary, although from a dump dating back to 2013. In this resource, contrarily to the one we built, relations are not typed with a sufficient granularity (it only distinguishes between a cognacy relation and a generic etymological origin relation).[20] Moreover, it relates non-glossed citation forms (rather than lexemes). Nevertheless, it is the only resource that is comparable with ours. We therefore evaluated our etymological relation database with respect to this resource.

EtymWordNet contains 473,433 direct yet non-typed etymological relations as well as 538,558 cognacy relations. As mentioned above, many cognation relations can be added based on other relations. The most interesting etymological information is provided by these other relations, which are unfortunately not distinguished within EtymWordNet. Another issue with EtymWordNet is that derivation and composition relations are not modelled in a satisfying way. For instance, (American) English *monophthongize* is the source of two independent etymological relations, one with *-ize* and the other one with *monophthong*.

To make the comparison possible, we had to transform our relations (excluding cognation relations) so that they follow the same model as EtymWordNet. Unsurprisingly, this slightly lowers the number of relations to 559,614. Among them, 464,542 (83%) are not found in EtymWordNet. Conversely, 378,361 relations are only found in EtymWordNet. But among these 378,361 relations, 333,369 (88%) relate forms from the same language: they are derivation or composition relations, extracted from other parts of the Wiktionary articles than the etymological part we exploited (especially the "derived terms" sections). Such relations are less interesting from an etymological point of view. Among the other missing relations, a large (yet hard to quantify) number of them are almost identical to relations that are included in our database, differing only by diacricts added in Wiktionary since 2013. Overall, this comparison shows that our database is significantly richer than EtymWordNet—and recall that our database relates (mostly glossed) lexemes with typed relations correctly representing inheritance, borrowings, derivation and composition, whereas EtymWordNet relates (non-glossed) citation forms with non typed relations (apart from the notion of cognacy relation).

7. Future work: improvement and use of our etymological database

The work presented here shall be improved in two ways. Firstly, patterns used for information extraction from etymological sections in Wiktionary articles can be extended,

[19] http://www1.icsi.berkeley.edu/~demelo/etymwn/
[20] We ignore relations of the type "orthographic variant".

improved, refined. Secondly, the lexeme merging algorithm can be enriched so as to merge lexemes which are not merged yet, mostly because variation of the following types:

- formal variations: differences in transcription or notation, form with or without stress information,[21] complete citation form vs. truncated citation form vs. principal parts;[22]
- variation in glosses[23] (for instance using WordNet or distributional similarity information).

The way we gloss lexemes that have no gloss in Wiktionary can also be improved, for instance by better taking advantage of their context of occurrences and by using external bilingual or multilingual resources.

A model of the phylogenetic relations between languages would also help replacing indirect relations with direct ones, either using simple heuristics or based on a (partial) model of phonetic (and maybe morphological) change. For instance, the relation Fr. *chapitre* 'chapter' < OFr. *chapitre* 'chapter' could be replaced by a relation MFr. *chapitre* 'chapter' < OFr. *chapitre* and a relation MFr. *chapitre* 'chapter' < OFr. *chapitre*, simply by knowing that, given our language inventory, the immediate ancestor of French is Middle French, whose immediate ancestor is Old French. It could help extending the lexicons for a number of intermediate languages with attested words, which could be validated using external lexical resources, or even unattested words.

Finally, it would be useful to extract the etymological information available in other Wiktionary editions, especially its French edition, the Wiktionnaire. Our databases are only affected by the language of the original information source at the level of glosses. We could automatically replace French glosses extracted from the Wiktionnaire by English glosses, for example by exploiting the English translations provided in the Wiktionnaire articles themselves.

In addition to lexicon extension for intermediate languages, as mentioned above, the resource presented in this article could be used as a starting point for research in computational historical linguistics, as suggested in the introduction. It may also be subjected to automated internal consistency checks, for example by automatically extracting phonetic laws and verifying their systematic applicability, modulo the analogical levelling phenomena. In the long term, this could also allows the construction or the automatic completion of large-coverage etymological dictionaries.

8. References

Bowers, J. & Romary, L. (2016). Deep encoding of etymological information in TEI. URL https://hal.inria.fr/hal-01296498. Working draft.

Buchi, E. (2016). Etymological dictionaries. In P. Durkin (ed.) *The Oxford Handbook of Lexicography*. Oxford University Press, pp. 338–349.

de Melo, G. (2014). Etymological Wordnet: Tracing the History of Words. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, LREC 2014. Reykjavik, Iceland, pp. 1048–1054.

[21] Gr. πλάξ 'flat stone' vs. Gr. πλαξ 'flat stone' are not merged yet.

[22] PIE *deh₂mo-* '(pas de glose)' and PIE *deh₂mos* 'people' are not merged yet.

[23] Fr. *aise* 'ease' and Fr. *aise* 'satisfaction, joy' are not merged yet, which prevents a further merging with Fr. *aise* '(pas de glose)' because of the resulting spurious ambiguity.

Salmon-Alt, S. (2006). Data structures for etymology: towards an etymological lexical network. *BULAG*, 31, pp. 1–12.