Construction automatique d'une base de données étymologiques à partir du wiktionary

Benoît Sagot

Construction automatique d'une base de données étymologiques à partir du *wiktionary*

Benoît Sagot
Inria, équipe ALMAnaCH, 2 rue Simone Iff, 75012 Paris, France
benoit.sagot@inria.fr

RÉSUMÉ

Les ressources lexicales électroniques ne contiennent quasiment jamais d'informations étymologiques. De telles informations, convenablement formalisées, permettraient pourtant de développer des outils automatiques au service de la linguistique historique et comparative, ainsi que d'améliorer significativement le traitement automatique de langues anciennes. Nous décrivons ici le processus que nous avons mis en œuvre pour extraire des données étymologiques à partir des notices étymologiques du *wiktionary*, rédigées en anglais. Nous avons ainsi produit une base multilingue de près d'un million de lexèmes et une base de plus d'un demi-million de relations étymologiques entre lexèmes.

ABSTRACT

Automatic construction of an etymological database using Wiktionary.

Electronic lexical resources almost never contain etymological information. The availability of such information, if properly formalised, would 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.

MOTS-CLÉS : Développement de ressources lexicales, étymologie, *wiktionary*.

KEYWORDS: Lexical resource development, etymology, Wiktionary.

1 Introduction

Les ressources lexicales électroniques utilisées en traitement automatique des langues et en linguistique informatique, dans leur très grande majorité, sont des ressources synchroniques : le plus souvent, elles fournissent des informations sur les propriétés flexionnelles, dérivationnelles, syntaxiques, sémantiques voire pragmatiques de leurs entrées. Le caractère formalisé en permet l'utilisation par des outils automatiques.

À l'opposé, les informations diachroniques, et notamment l'étymologie, ne font qu'exceptionnellement l'objet de ressources électroniques et restent ainsi l'apanage des dictionnaires imprimés ou publiés en ligne (comme par exemple le TLFi). Les rares ressources existantes, et notamment la

base de données *The Tower of Babel* [1] ou le projet *PIElexicon* [2], s'appuient souvent sur des principes comparatifs et étymologiques qui sont au mieux obsolètes et non consensuels [3] et au pire complètement fantaisistes [4]. Seul l'EtymWordNet (de Melo, 2014), sur lequel nous reviendrons, fait figure d'exception malgré ses sévères limitations.

La disponibilité de bases de données étymologiques formalisées, riches et de grande couverture serait pourtant de nature à permettre des avancées importantes en linguistique diachronique, historique et comparative. En effet, la modélisation de l'évolution des langues et la reconstruction de proto-langues, ancêtres de plusieurs langues attestées, repose sur la prise en compte de données lexicales très conséquentes, couvrant souvent des dizaines voire des centaines de langues. Pour certaines familles comme les langues indo-européennes ou les langues sémitiques, près de deux siècles de recherches minutieuses ont permis d'avoir une assez bonne compréhension de la diachronie lexicale. Toutefois, même pour ces deux familles, et *a fortiori* pour les autres, de nombreuses zones d'ombre persistent.

Développer des moyens automatiques pour explorer les correspondances formelles et sémantiques potentielles entre mots de langues différentes et pour modéliser leur évolution diachronique permettrait de renouveler les sous-disciplines linguistiques concernées, tout en posant des problèmes algorithmiques difficiles. Cela contribuerait aussi au développement de ressources et d'outils pour le traitement automatique de documents anciens qui reflètent un état de la langue significativement différent de ce que les outils habituels peuvent traiter (par exemple, des documents en ancien ou moyen français, que des outils de traitement automatique du français contemporain ne peuvent traiter correctement). De telles entreprises ne peuvent naturellement que s'appuyer sur le résultat des travaux étymologiques antérieurs, qu'il convient donc de représenter sous la forme de ressources lexicales électroniques formalisées.

Pour cela, il est nécessaire d'identifier une source exploitable d'informations étymologiques, d'en extraire automatiquement ces informations et de les représenter sous une forme structurée voire normalisée. C'est un tel travail que nous décrivons dans cet article. Nous nous appuyons sur l'édition anglaise du *wiktionary* [5], dictionnaire collaboratif en ligne dont la syntaxe est semi-structurée et qui inclut des informations étymologiques assez riches et d'une fiabilité raisonnable, naturellement rédigées en anglais [6]. Cet article est structuré comme suit. Après un bref aperçu des travaux antérieurs ayant un lien avec le nôtre (section 2), et un survol des différents types de relations étymologiques entre lexèmes (section 3), nous décrivons les notices étymologiques que l'on trouve dans le *wiktionary* et la façon dont nous les avons extraites et partiellement structurées (section 4). La section 5 est consacrée au processus de construction d'une base de lexèmes et d'une base de relations étymologiques entre lexèmes. La section 6 fournit un certain nombre d'informations quantitatives sur ces bases, les formats dans lesquels nous les avons exportées, ainsi qu'une évaluation manuelle de leur qualité. Nous discuterons enfin à la section 7 des suites à donner à ce travail, et notamment de certaines de ses applications immédiates.

1. `http://starling.rinet.ru/babel.php?lan=en`
2. `http://pielexicon.hum.helsinki.fi`
3. La base de données indo-européenne de *The Tower of Babel* s'appuie sur le dictionnaire de Pokorny, aujourd'hui obsolète. De plus, les auteurs de cette base défendent des hypothèses peu conventionnelles sur des liens de parenté entre familles de langues traditionnelles, hypothèses généralement rejetées mais qui influencent certaines de leurs propositions étymologiques.
4. C'est le cas pour le *PIElexicon*, bien que la justification de cette affirmation dépasse le cadre de cet article.
5. `https://en.wiktionary.org/`
6. L'édition anglaise du *wiktionary* est, de loin, celle qui contient les informations étymologiques les plus nombreuses, les plus complètes et les plus fiables. Pour le français, l'édition française (le wiktionnaire) est également une source d'informations étonnamment fiable, mais que nous n'avons pas exploitée dans ce travail ;

Les deux bases ainsi construites sont librement disponibles sous licence LGPL-LR.

2 Travaux antérieurs

Les travaux antérieurs en lien avec le travail présenté ici peuvent être classés en trois catégories : les travaux concernant la normalisation des informations étymologiques, les bases de données existantes et, pour finir, l'EtymWordNet déjà mentionné.

Les informations étymologiques n'étant qu'exceptionnellement prises en compte dans les ressources lexicales électroniques, leur représentation structurée ne fait pas encore l'objet de recommandations quant à leur normalisation. À cet égard, le document de travail publié par Bowers & Romary (2016) reflète l'état des recherches. Il s'appuie sur plusieurs initiatives antérieures, et notamment sur les travaux de Salmon-Alt (2006). Il propose un ensemble de principes généraux pour la représentation des informations étymologiques dans les dictionnaires électroniques encodés en TEI. Il s'appuie sur une typologie relativement large des phénomènes sous-jacents, qui couvre l'héritage standard (ce que les étymologistes qualifient de *recto itinere*, 'en droite ligne'), l'emprunt, la métaphore, la métonymie, la composition et la grammaticalisation. Certains de ces mécanismes ne sont pas étymologiques à proprement parler mais ressortent plutôt de mécanismes de création lexicale. Nous reviendrons à la section 6 sur certaines limites de l'état actuel de cette proposition.

Peu de dictionnaires électroniques librement disponibles font usage de représentations structurées des informations étymologiques. Nous avons déjà mentionné *The Tower of Babel* et le *PIElexicon*. Un autre exemple est le *Germanic Lexicon Project*[7] de S. Crist, dont le format de représentation peut également être considéré comme un précédesseur des propositions de Bowers & Romary (2016). Toutefois, les différents dictionnaires libres de droits distribuées dans ce cadre ne sont que faiblement structurés : l'extraction systématique de relations étymologiques serait une tâche non triviale. Ce n'est pas le cas de la *World Loanword Database*, qui, pour 1 460 sens sélectionnés avec soin, fournit un ou plusieurs lexèmes dans 41 langues, chacun étant associé à un niveau de probabilité de résulter d'un emprunt ainsi que des lexèmes sources possibles. Mais l'inventaire des 41 langues couvertes reflète le positionnement typologique et non étymologique du projet. En tout état de cause, on est loin d'une ressource à large couverture, et, bien sûr, seuls les mécanismes d'emprunt sont couverts, à l'exclusion de tout autre mécanisme étymologique.

Plus proche de notre travail, de Melo (2014) a rendu disponible l'EtymWordNet, qu'il a, comme nous, extrait automatiquement à partir du *wiktionary* anglais (quoique dans une version de trois ans plus ancienne). Toutefois, et malgré une large couverture, l'EtymWordNet n'est pas exploitable tel quel pour l'informatisation de la linguistique comparée et historique en raison de deux limitations fondamentales : les mécanismes en jeu ne sont pas distingués (par exemple, aucune distinction entre héritage, emprunt et dérivation morphologique) et, plus grave encore, les unités manipulées sont des lemmes et non des lexèmes : les sens sont ignorés.

Nous n'avons pas connaissance de travaux antérieurs ayant eu pour résultat une base étymologique électronique formalisée à large couverture qui manipule des lexèmes, comme il est nécessaire en lexicologie étymologique (cf. section 3), et qui distingue les mécanismes étymologiques entre eux. C'est là l'objectif de notre travail.

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

3 Mécanismes étymologiques et de création lexicale

L'extraction et la formalisation d'informations étymologiques nécessite de disposer d'un modèle, même simple, de ce type d'informations. La première question qui se pose est celle de l'unité de base. Comme rappelé par Buchi (2016, p. 346), il ne saurait s'agir que du lexème, c'est-à-dire dans notre cas d'une forme de citation, d'un identifiant de langue et d'une glose en langue anglaise [8]. Naturellement, le passage d'un lexème à un autre peut modifier chacun de ces trois éléments : la langue (changement diachronique dans le cas de l'héritage, synchronique dans le cas de l'emprunt), la forme de citation (changements phonétiques mais également morphologiques) et le sens (évolutions sémantiques).

La seconde question qui se pose est celle de la nature de la relation étymologique entre lexèmes. Suivant à nouveau Buchi (2016, p. 346–7), une relation étymologique élémentaire doit se faire entre lexèmes immédiatement reliés : il s'agit d'un lien *direct*. Dans le cas d'un héritage *recto itinere*, et dès lors que l'on se dote d'un inventaire (nécessairement arbitraire) d'identifiants de langues, une relation élémentaire doit donc impliquer un lexème d'une langue à un ou plusieurs lexèmes de l'état immédiatement antérieur de cette même langue [9]. Dans le cas d'un emprunt, une relation directe associe simplement le lexème cible au lexème source [10]. En effet, utiliser autant que possible des relations directes est la seule façon de pouvoir spécifier la nature du procédé étymologique en cause [11].

La troisième question qui se pose est celle des différents types de mécanismes étymologiques. Bien que nous n'en traiterons pas tous les cas de figure, nous utiliserons la typologie suivante :

- Héritage (avec évolution phonétique si besoin ; avec ou sans évolution du sens et/ou de la morphologie) ; comme il est d'usage, nous noterons cette relation comme suit : *lexème cible* $<$ *lexème source* ;
- Emprunt ; nous noterons cette relation comme suit : *lexème cible* $\leftarrow$ *lexème source* [12] ;
- Création lexicale
  - Dérivation morphologique
    - Dérivation suffixale ; elle sera notée ainsi : *lexème cible* $<_s$ *base* $+$ *suffixe* ;
    - Dérivation préfixale ; elle sera notée ainsi : *lexème cible* $<_p$ *préfixe* $+$ *base* ;
    - Autres cas (y compris les dérivés construits par analogie) ; ils seront notés ainsi : *lexème cible* $<_d$ *élément* $+ \ldots +$ *élément* ;
  - Composition morphologique, notée ainsi : *lexème cible* $<_c$ *composant* $+ \ldots +$ *composant* ;
  - Création d'un mot-valise, cas non traité dans cet article ;
  - Troncation et autres phénomènes similaires, cas non traités dans cet article.

À cet inventaire nous rajouterons une relation spéciale de cognation, qui pourra relier deux lexèmes (d'une même langue ou de deux langues différentes) qui ont une étymologie au moins partiellement commune (en général, au moins une même « racine »). Elle sera notée *lexème*$_1$ // *lexème*$_2$.

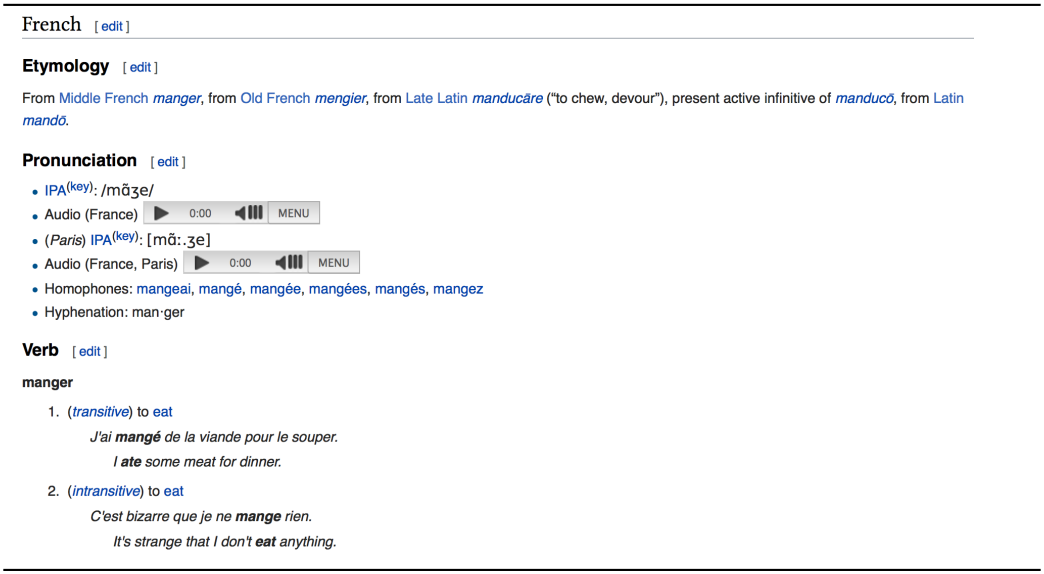
8. Cela recouvre aussi le cas des noms de lieux, de personnes, de peuples, et d'autres types de noms propres.

9. Fr. *manger* $<$ Moy. Fr. *manger* est ainsi un lien direct, contrairement à Fr. *manger* $<$ Anc. Fr. *mengier* voire Fr. *manger* $<$ Lat. Tard. *manducāre* qui sont des liens indirects.

10. Ainsi, des liens comme Fr. *abricot* $<$ Esp. *albaricoque* ou Fr. *abricot* $<$ Port. *albricoque* sont de possibles liens directs (les deux sont plausibles). Les mots espagnol et portugais sont empruntés à l'arabe *al-barqūq*, lui-même emprunté au grec médiéval βερικόκκια 'abricotier', dérivé du grec ancien πραικόκιον 'abricot' qui, lui-même, est un emprunt au Lat. *praecoquum* '(fruit) précoce'. Ainsi, un lien tel que Fr. *abricot* $<$ Lat. *praecoquum* serait correct mais pas direct.

11. Reprenant l'exemple de la note précédente, il serait bien délicat de qualifier le lien entre Lat. *praecoquum* et Fr. *abricot*, qui recouvre de multiples étapes de natures différentes.

12. Nous incluons dans les emprunts les cas d'emprunts savants comme Fr. *oculaire* 'ocular' $\leftarrow$ Lat. *ocularis* 'ocular'.



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): [mɑ̃ːʒe]
- 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 – Extrait d'une entrée du *wiktionary*.

4 Extraction et structuration des notices étymologiques

4.1 Informations étymologiques dans le *wiktionary*

Le *wiktionary*, dont nous utiliserons ici l'édition anglaise, est un dictionnaire multilingue collaboratif. Il est organisé en articles qui peuvent contenir des entrées lexicales pouvant relever de plusieurs langues, mais qui ont pour point commun d'avoir comme forme de citation le titre de l'article. Il peut naturellement y avoir dans un article plusieurs entrées lexicales distinctes pour une langue donnée, qui peuvent différer par la partie du discours, par l'étymologie ou par simple homonymie.

Nous avons utilisé le *dump* en date du 01/01/2017. Il contient près de 5 millions et demi d'articles dont plus de 40 000 sont des pages de redirection. Ces articles contiennent au total 894 453 notices étymologiques ou sous-titres distinguant des étymologies distinctes [13]. Ce *dump* est dans un format semi-structuré : la structuration en articles est encodée en XML, y compris des métadonnées sur chaque article ; le contenu de chacun des articles est codé au moyen de la syntaxe dite « wiki ». Il s'agit, pour simplifier, de texte brut complété par des marqueurs typographiques (différents niveaux de titres, listes, etc.) et de *templates* permettant de coder certaines informations de façon systématique. Par exemple, la template `link` (ou `l`) permet d'intégrer à l'article une forme qui est un lien vers l'article dont elle est le titre. Ainsi, `{{link|fr|chaise||chair|g=f}}` sera rendu sur le site de wikipedia par chaise *f* ("chair"), où le genre féminin est indiqué (`g=f`) et où le mot chaise est un hyperlien vers la partie de l'article correspondant concernant les lexèmes français (`fr`) [14].

13. La présence d'un tel sous-titre n'indique pas nécessairement la présence d'une notice étymologique (pour un exemple, cf. l'article *scritch* sur le *wiktionary*).

14. L'inventaire de langues utilisé par *wiktionary* s'appuie sur les normes ISO-639-1 à ISO-639-3, tout en les étendant en cas de besoin. Pour plus de détails, cf. `https://en.wiktionary.org/wiki/Wiktionary:Languages`. Nous nous sommes également dotés, par l'intermédiaire de la correspondance entre codes langues et noms de langues, d'un système

```
==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 – Code source correspondant à l'extrait d'article de la figure 1 (le code concernant la prononciation a été omis).

La figure 1 montre un extrait de l'article « *manger* » du *wiktionary*. Le code source correspondant est donné à la figure 2.

Enfin, des sections « Descendants » sont parfois présentes. Elles listent des descendants du lexème courant, sans préciser la nature du lien étymologique concerné (héritage, emprunt).

4.2 Extraction et structuration

Nous avons converti le *dump* du *wiktionary* en un fichier XML au moyen de simples cascades d'expressions régulières [15]. Ce fichier XML est une liste d'entrées lexicales qui correspondent approximativement à des lexèmes. Il ne contient que des entrées pour lesquelles le *wiktionary* fournit des informations étymologiques dans une section dédiée. Le nombre total d'entrées obtenues est de 831 988. Chacune d'elles reprend le contenu de cette section étymologique et l'inclut dans une balise `<etymology/>`. Cette notice est préalablement reformatée de telle sorte que les formes qui y sont citées, notamment mais pas seulement celles qui le sont au moyen de *templates*, soient représentées par un élément XML `<form/>`. Les situations impliquant simultanément plusieurs `<form/>` (dérivation affixale, composition) sont harmonisées à l'aide du symbole « + » (cf. plus haut). Les cas où plusieurs formes alternantes (variantes, parties principales…) sont listées sont également harmonisées à l'aide du symbole « ∼ ». Ces étapes de normalisation, en apparence simples, sont rendues complexes de par la variété des situations, la richesse des *templates* utilisables et la multiplicité des moyens utilisés par les contributeurs du *wiktionary* pour représenter l'information.

Si une section listant les descendants est présente, ils sont convertis en éléments `<form/>` et sont inclus dans un élément `<descendants/>` à l'intérieur de l'`<etymology/>`.

d'abréviation automatique déterministe des noms de langues, ainsi que d'un système d'identification des (codes) langues à partir de leur nom ou de leurs abréviations usuelles telles qu'utilisées dans les articles. Ainsi, « OFr. », « Old Fr. » ou « Old French » sont correctement interprétés comme reflétant la langue de code `fro`, code qui peut ensuite être transformé en son abréviation standard en anglais, à savoir « OFr. ».

15. Ce format XML est un format de travail. Il n'a pas vocation à ce stade à être adapté pour le rendre compatible avec la TEI. Nous reviendrons à la section 6.2 sur l'export en TEI des relations étymologiques seulement, à l'exclusion du reste des notices.

```
<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 – Résultat de notre processus de structuration à partir du code source indiqué à la figure 2.

Toutes les formes citées ailleurs que dans la section étymologique ou la section listant des descendants sont également extraites et listées dans une section `<forms/>`. En effet, ces formes pourront s'avérer utiles par la suite, notamment celles qui sont munies d'une glose.

Lorsque cela est possible, le lexème courant est muni d'une glose. S'il s'agit d'un lexème anglais, il est considéré comme étant sa propre glose. Dans les autres cas, nous tentons d'extraire une ou plusieurs gloses à partir des définitions fournies dans l'article.

À partir du code source concernant le verbe français *manger* indiqué à la figure 2, notre processus de structuration produit ainsi l'entrée montrée à la figure 3.

5 Construction de la base de relations étymologiques

Le résultat du processus de structuration du *dump* du *wiktionary* décrit à la section précédente est plus exploitable que le *dump* lui-même, mais pose néanmoins un certain nombre de difficultés. La plus importante est naturellement qu'à l'exception des éléments `<form/>`, les informations étymologiques sont en anglais, sans aucune formalisation. Une autre limite est que, d'un article à l'autre, un même lexème peut être glosé de différentes façons, et parfois ne pas l'être du tout.

Pour tenter de remédier à ces difficultés, nous avons procédé en plusieurs étapes. Tout d'abord, nous avons défini un certain nombre de motifs permettant d'inférer la glose d'une forme, lorsqu'elle n'a pas été explicitée, à partir du contexte [16]. L'élément `<form/>` est alors complété en conséquence.

Nous parcourons alors l'ensemble des sections étymologiques et créons des triplets de la forme (lexème cible, lexème ou séquence de lexèmes source, type de la relation) à l'aide d'une cascade de motifs réguliers. Au cours de ce processus, chaque `<form/>` est associé à un identifiant entier, et chaque triplet associe deux de ces identifiants et un type de relation. Les composés et dérivés font l'objet d'un traitement particulier, au moyen d'identifiants négatifs. Supposons par exemple que l'on ait identifié que *bêtement* $<_s$ *bête* $+$ *-ment* dans l'entrée de Fr. *bêtement*, dont les définitions en anglais dans le *wiktionary* ont été correctement identifiées et ont permis d'extraire la glose 'stupidly, idiotically'. S'il n'existe pas déjà, on attribue un identifiant entier positif i non encore utilisé au

16. Pour reprendre l'exemple de *manger*, l'indication « From Middle French *manger*, from Old French *mengier*… », bien que ne contenant aucune glose, rend possible l'attribution de la glose du lexème vedette *manger*, à savoir 'to eat', au moyen français MFr. *manger* et à l'ancien français OFr. *mangier*.

lexème Fr. *bêtement* ‘stupidly, idiotically’. On fait de même avec Fr. *bête* ‘(pas de glose)’ et Fr. *-ment* ‘(pas de glose)’. S’il n’existe pas déjà, on associe alors un identifiant entier négatif j non encore utilisé au couple formé des identifiants respectifs de Fr. *bête* ‘(pas de glose)’ et de Fr. *-ment* ‘(pas de glose)’. On peut alors créer le triplet $(i, j, \mathrm{der(s)})$, où « $\mathrm{der(s)}$ » indique le type de relation, à savoir une dérivation suffixale.

Ceci étant fait, il nous faut fusionner, autant que possible, les lexèmes synonymes : par exemple, si d’autres notices étymologiques ont conduit à attribuer des identifiants (entiers positifs) à des lexèmes tels que Fr. *bêtement* ‘(pas de glose)’ ou Fr. *bêtement* ‘stupidly, foolishly’, il convient de les fusionner tous deux avec le lexème Fr. *bêtement* ‘stupidly, idiotically’ mentionné précédemment. Pour ce faire, nous itérons jusqu’à stabilité la séquence d’étapes suivantes :

- si un lexème sans glose a la même langue et la même forme de citation qu’un unique autre lexème (nécessairement glosé), alors ces deux lexèmes sont fusionnés ;
- si deux lexèmes glosés ont la même langue, la même forme de citation, et au moins une glose en commun (cf. ‘stupidly, foolishly’ vs. ‘stupidly, idiotically’), alors ils sont fusionnés (résultant dans cet exemple en une glose complète ‘stupidly, foolishly, idiotically’) ;
- toutes les relations, ainsi que les définitions des identifiants négatifs représentant les composés et les dérivés, sont alors mis à jour en conséquence des fusions de lexèmes ainsi réalisées.

Pour obtenir, dans la mesure du possible, des relations étymologiques directes, nous supprimons toute relation entre deux lexèmes *lexème*$_1$ et *lexème*$_3$ telle qu’il existe également une relation entre *lexème*$_1$ et un lexème intermédiaire *lexème*$_2$ et une relation entre ce *lexème*$_2$ et *lexème*$_3$ [17].

Enfin, le type de certaines relations est corrigé, afin de préciser autant que possible les cas où l’on est en présence d’emprunts ou de dérivation morphologique et non d’héritage, qui reste néanmoins la relation par défaut.

Le résultat de ce processus d’extraction est double : un ensemble de lexèmes, dont certains seulement sont glosés, et un ensemble de relations impliquant un lexème cible, un ou plusieurs lexèmes sources (plusieurs en cas de composition ou dérivation affixale) et un type de relation. En voici quelques exemples réels concernant des lexèmes français :

- Fr. *gobelet* ‘goblet’ $<$ OFr. *gobel* ‘goblet ; cup ; beaker ; tumbler’
- Fr. *maudire* ‘to curse’ $<$ OFr. *maudire* $\sim$ *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’ $\leftarrow$ Port. *acajú* ‘cashew’
- Fr. *car* ‘car ; coach’ $\leftarrow$ E *car*

6 Résultats et évaluation

6.1 Données quantitatives

L’étape initiale d’extraction décrite au début de la section 5 a produit près d’1,2 million de lexèmes, 62 056 séquences de lexèmes et 548 935 relations étymologiques entre deux lexèmes ou entre un lexème et une séquence de lexèmes [18]. Quelques dizaines d’itérations de l’algorithme de fusion conduisent à la fusion de 199 185 lexèmes et de 289 séquences de lexèmes d’où au final 975 473

17. Le même mécanisme s’applique lorsque *lexème*$_3$ n’est pas un lexème unique mais une séquence de lexèmes.
18. Dans ces nombres comme dans les suivants, les lexèmes qui ne font partie d’aucune relation ne sont pas comptabilisés.

lexèmes distincts, 61 809 séquences de lexèmes distinctes et 519 348 relations distinctes. L'élimination de 5 149 relations non-directes conduit à un nombre final de relations égal à 514 199.

Les lexèmes obtenus relèvent de 2311 langues distinctes, parmi lesquelles les mieux représentées sont, dans l'ordre, l'anglais (257 978 lexèmes), le latin (65 981), le français (32 044), l'italien (28 028) et le grec ancien (21 077). Parmi ces lexèmes, 659 567 soit 68% disposent d'une glose.

Parmi les 514 199 relations, 452 041 sont entre deux lexèmes, les autres reliant un lexème et une séquence de lexèmes. Les relations de cognation sont au nombre de 90 511, les 423 673 autres étant des relations directes. Enfin, 318 883 des relations obtenues n'impliquent que des lexèmes glosés.

Nous aurions pu facilement augmenter de façon importante le nombre de relations de cognation, en ajoutant des relations entre lexèmes ayant un ancêtre (direct ou indirect) commun dans notre base.

6.2 Inférence et export en TEI des chaînes étymologiques

Nous avons développé un module d'export de notre base de relations étymologiques dans le format TEI proposé par Bowers & Romary (2016). Dans ce format, les relations directes peuvent être exportées sous forme d'éléments `<etym/>` simples, associées au type de relation concerné.

Cependant, il peut être intéressant de disposer pour un lexème non seulement de son étymon direct mais également de son histoire étymologique aussi complète que possible. C'est d'ailleurs ce que fournissent un nombre significatif de notices étymologiques du *wiktionary*, comme illustré aux figures 1 et 2. Pour permettre de rétablir ces chaînes étymologiques à partir de notre base de relations, il suffit de chercher récursivement les relations impliquant chaque étymon considéré : à partir de Fr. *manger* 'to eat' $<$ MFr. *manger* 'to eat' et de MFr. *manger* 'to eat' $<$ OFr. *mengier* 'to eat' on peut reconstruire la chaîne Fr. *manger* 'to eat' $<$ MFr. *manger* 'to eat' $<$ OFr. *mengier*. C'est l'approche que nous avons suivie.

Nous avons dû étendre la proposition de Bowers & Romary (2016) dans quatre directions, ce qui pourrait servir d'inspiration pour son amélioration :

- Cette proposition ne couvre pas la relation de cognation. Nous proposons pour cela l'ajout d'un type supplémentaire (*type="cognate"*) à l'élément `<etym/>`.
- Elle ne permet pas non plus de renvoyer à une autre entrée lexicale où des informations étymologiques pertinentes sont fournies. Nous proposons pour cela l'ajout d'un type supplémentaire (*type="reference"*) à l'élément `<etym/>`, à l'intérieur duquel une référence directe à l'entrée concernée peut être incluse au moyen d'un élément `<xr/>` (élément TEI utilisé pour indiquer une référence croisée).
- Bowers & Romary (2016) ne proposent pas de moyen pour encoder les chaînes étymologiques. On peut simplement utiliser un élément `<etym/>` spécial qui, par un attribut, indique qu'il encapsule une séquence de relations étymologiques, lesquelles sont représentées par des éléments `<etym/>` spécifiques à l'intérieur de l'`<etym/>` global.
- Dans leur document, Bowers & Romary (2016) ne permettent pas d'indiquer des hypothèses alternatives, cas fréquent dans nos données. Nous utilisons là encore un élément `<etym/>` spécial qui, par un attribut, indique qu'il encapsule une alternative entre deux chaînes étymologiques, lesquelles sont représentées par des éléments `<etym/>` distincts.

Dans les deux derniers cas, le caractère récursif des éléments `<etym/>` permet toutes les combinaisons possibles, comme par exemple une chaîne étymologique commençant par deux étapes « certaines » puis se poursuivant par une alternative entre deux sous-chaînes étymologiques différentes.

```
<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 – Exemple d'entrée exportée au format TEI (les relations de cognations sont ignorées).

6.3 Évaluation manuelle

L'évaluation de notre base de relations étymologiques peut s'envisager sous quatre angles :

1. Quelle est la qualité des informations étymologiques présentes dans le *wiktionary* ?
2. Quelles sont les erreurs liées au processus de structuration des notices étymologiques ?
3. Quelles sont les erreurs introduites par les algorithmes d'inférence de gloses et de fusion entre lexèmes ? À l'inverse, quel est le degré de couverture de ces algorithmes ?
4. Quelles sont les erreurs qui résultent d'une interprétation systématique des relations non-typées (c'est-à-dire qui ne sont pas spécifiées comme de type emprunt ou de type dérivation suffixale, par exemple) comme étant des relations d'héritage ?

Répondre précisément à la première question est délicat et dépasse le cadre de cet article. L'étude aléatoire de quelques dizaines de notices étymologiques montre que les informations qu'elles contiennent sont généralement fiables. Seules les étymons proto-indo-européens font parfois usage de notations reflétant un état dépassé des connaissances. On peut toutefois retenir que les informations étymologiques du *wiktionary* sont très majoritairement fiables et reflètent souvent les publications spécialisées les plus récentes ou les plus consensuelles, qui sont du reste régulièrement citées en référence.

La précision et le rappel de nos algorithmes d'inférence de gloses et de fusion de lexèmes sont plus faciles à estimer. Nous nous sommes tout d'abord concentrés sur le rappel de l'algorithme de fusion de lexèmes. Pour cela, nous avons extrait aléatoirement 50 couples (langue, forme) parmi les 124 775 (sur 941 757) qui correspondent à plusieurs entrées. Nous avons ensuite extrait toutes les entrées correspondant à ces 50 couples, et avons estimé manuellement la pertinence de leur co-existence. Dans la quasi-totalité des cas, des fusions supplémentaires auraient dû avoir lieu, mais notre algorithme ne le permettait pas. C'est donc une piste d'amélioration évidente. À l'inverse, pour estimer la précision de l'algorithme de fusion de lexèmes ainsi que celle de notre algorithme

d'inférence de glose, nous avons extrait aléatoirement 100 formes glosées et avons vérifié l'exactitude et la cohérence de leurs gloses. Sur ces 100 formes nous avons identifié 2 erreurs d'extraction, dues toutes deux à une utilisation inhabituelle de la syntaxe « wiki » par les contributeurs, une erreur partielle (certaines des gloses fournies sont correctes, l'une d'entre eux est du code wiki facile à ignorer), une transcription prise pour une glose et une définition (correcte) prise pour une glose. Tous les autres cas étaient valides. Il y a donc quelques erreurs, mais elles ne proviennent (quasiment) jamais de nos algorithmes d'inférence de glose et de fusion de lexèmes. C'est plutôt au niveau du processus de structuration des notices étymologiques que des améliorations ponctuelles pourraient être réalisées.

Enfin, nous avons évalué les relations proprement dites sur un échantillon aléatoire de 100 relations étymologiques. Parmi elles, 78 sont correctes, 18 sont de type « héritage » alors qu'elles représentent des emprunts, 3 ont d'autres types d'erreurs de typage, et une seule est erronée en raison d'un problème dans l'extraction de la forme source. Les 18 erreurs où le type « héritage » est extrait à la place du type « emprunt » sont le reflet de ce que la relation par défaut, en l'absence d'indication explicite, est considérée comme étant l'héritage. Une description plus fine des relations entre langues permettrait de corriger facilement ces erreurs. C'est quelque chose qui sera fait dans un proche avenir.

6.4 Comparaison avec l'*EtymWordNet*

L'EtymWordNet (de Melo, 2014), librement téléchargeable sans licence explicite [19], est une base de données étymologiques extraite du *wiktionary*, quoique dans une version datant de 2013. Dans cette ressource, contrairement à celle que nous avons construite, les relations ne sont pas typées avec une granularité suffisante (seule l'origine étymologique et la relation de cognation sont distinguées [20]) et relient des formes de citations non glosées (par opposition à des lexèmes). Il s'agit toutefois de la seule ressource comparable avec la nôtre. Nous avons donc cherché à évaluer les relations que nous avons extraites par rapport à cette ressource existante.

L'EtymWordNet contient 473 433 relations étymologiques directes mais non typées ainsi que 538 558 relations de cognation. Comme évoqué plus haut, de nombreuses relations de cognation peuvent être facilement ajoutées à partir des autres relations. L'information étymologique fondamentale est celle fournie par les autres types de relations, malheureusement non distingués dans l'EtymWordNet. Une autre difficulté avec l'EtymWordNet est que les relations de dérivation et de composition ne sont pas modélisées correctement. Par exemple, l'anglais *monophthongize* est la source de deux relations étymologiques indépendantes, l'une avec *-ize* et l'autre avec *monophthong*.

Pour rendre la comparaison possible, nous avons donc transformé nos relations (celles de cognations exclues) en suivant la même représentation qu'EtymWordNet. Sans surprise, le nombre de nos relations diminue alors légèrement, pour atteindre un nombre de 559 614. Parmi elles, 464 542 (soit 83%) sont absentes de l'EtymWordNet alors que 95 072 y sont présentes. À l'inverse, 378 361 relations ne sont présentes que dans l'EtymWordNet. Mais parmi ces dernières, 333 369 (soit 88%) relient deux formes d'une même langue : il s'agit de relations de dérivation ou de composition, extraites de parties des articles que nous n'avons pas exploitées (« Derived terms » notamment). Mais ces relations sont moins intéressantes sur le plan étymologique. Parmi les autres relations manquantes, un certain nombre, que nous n'avons pas pu quantifier exactement, sont quasiment identiques à des relations présentes dans notre ressource et n'en diffèrent que par un diacritique rajouté dans le

19. `http://www1.icsi.berkeley.edu/~demelo/etymwn/`
20. Nous laissons de côté les liens de type « variante orthographique ».

wiktionary depuis 2013. Mais globalement, on peut conclure de cette comparaison que notre base est significativement plus riche qu'EtymWordNet — sans même rappeler qu'elle relie des lexèmes, majoritairement glosés, que les composés et dérivés y sont représentés explicitement et que les types de relations étymologiques y sont précisés.

7 Perspectives d'amélioration et d'utilisation

Le travail présenté ici devra être amélioré de plusieurs façons. Tout d'abord, les motifs de structuration des notices étymologiques puis d'extraction des lexèmes et des relations peuvent être étendus, améliorés, raffinés. Ensuite, l'algorithme de fusion des formes peut être enrichi afin d'unifier des lexèmes qui ne le sont pas encore en raison de variations qui peuvent être de deux types :

— variations formelles : différences de transcription ou de notation, forme accentuée vs. non accentuée [21], forme de citation complète vs. forme conventionnelle tronquée vs. séquence de parties principales [22] ;
— variations dans les gloses [23] (par exemple au moyen de WordNet ou de bases de similarités distributionnelles).

L'attribution de gloses aux lexèmes qui n'en ont pas encore pourra être améliorée, soit par une étude plus fine du contexte, soit par l'utilisation de ressources lexicales bilingues ou multilingues externes.

Une modélisation de l'arbre phylogénétique des langues utilisées permettrait également de remplacer certaines relations indirectes par des relations directes, soit au moyen d'heuristiques simples soit grâce à une modélisation au moins partielle des changements phonétiques voire morphologiques intervenus entre temps. Pour illustrer ceci sur un cas particulièrement simple, on pourrait remplacer la relation Fr. *chapitre* 'chapter' < OFr. *chapitre* 'chapter' par une relation Fr. *chapitre* 'chapter' < MFr. *chapitre* et une relation MFr. *chapitre* 'chapter' < OFr. *chapitre*. Cela permettrait d'étendre les lexiques de certaines langues intermédiaires au moyen de mots qui peuvent être attestés, et donc validés par la consultation de ressources externes si elles existent, ou ne pas l'être encore.

Enfin, il serait utile d'extraire les notices étymologiques disponibles dans certaines autres éditions de *wiktionary*, et notamment son édition française, le wiktionnaire. Les représentations formalisées utilisées par nos bases de lexèmes et de relations étymologiques ne reflètent la langue de la ressource d'origine que par la langue des gloses. On peut imaginer remplacer automatiquement des gloses en français extraite du wiktionnaire par des gloses en anglais, par exemple en exploitant les traductions anglaises fournies dans les articles du wiktionnaire eux-mêmes.

Outre la complétion de lexiques pour des langues intermédiaires, la ressource dont nous avons présenté la construction dans cet article pourra servir de point de départ à différents types de travaux en linguistique historique computationnelle, comme évoqué dans l'introduction. Elle pourra également faire l'objet de vérifications automatisées de cohérence interne, par exemple par l'extraction automatique de lois phonétiques et la vérification de leur applicabilité systématique, modulo les phénomènes de réfection analogiques. Elle pourrait permettre également, à terme, la construction ou la complétion automatique de dictionnaires étymologiques à large couverture.

21. Gr. πλάξ 'flat stone' et Gr. πλαξ 'flat stone' ne sont pas fusionnés.
22. Ainsi PIE *deh₂mo-* '(pas de glose)' et PIE *deh₂mos* 'people' ne sont pas fusionnés.
23. Ainsi, pour l'instant, Fr. *aise* 'ease' et Fr. *aise* 'satisfaction, joy' ne sont pas fusionnés, et empêchent la fusion avec le lexème Fr. *aise* '(pas de glose)' en raison de l'ambiguïté apparente qui en résulte.

Références

BOWERS J. & ROMARY L. (2016). Deep encoding of etymological information in TEI. document de travail.

BUCHI E. (2016). Etymological dictionaries. In P. DURKIN, Ed., *The Oxford Handbook of Lexicography*, p. 338–349. Oxford University Press.

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, p. 1048–1054, Reykjavik, Islande.

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