Etymological Wordnet: Tracing The History of Words

Gerard de Melo
IIIS, Tsinghua University
Beijing, P.R. China
gerard@demelo.org

Abstract

Research on the history of words has led to remarkable insights about language and also about the history of human civilization more generally. This paper presents the Etymological Wordnet, the first database that aims at making word origin information available as a large, machine-readable network of words in many languages. The information in this resource is obtained from Wiktionary. Extracting a network of etymological information from Wiktionary requires significant effort, as much of the etymological information is only given in prose. We rely on custom pattern matching techniques and mine a large network with over 500,000 word origin links as well as over 2 million derivational/compositional links.

Keywords: etymology, historical linguistics, multilingual resources

1. Introduction

Investigating the origins of words can lead to remarkable insights about the cultural background that has shaped the semantics of our modern vocabulary. As a matter of fact, research in comparative and historical linguistics has not only produced numerous invaluable findings about the history of words and languages but also about the history of humanity and the migration patterns that have shaped our world.

Often, however, research in this area is concerned with very specific languages and time periods rather than aiming at large-scale data aggregation across many language families. Additionally, etymological relationships are typically described in prose. While the background information that such prosaic form can provide is undoubtedly significant, this makes it harder for machines to observe the essential connections between words. For these reasons, there has not been any machine-readable resource that aggregates large numbers of etymological relationships across thousands of words in hundreds of languages.

In this paper, we present the Etymological Wordnet, a lexical resource that attempts to make a major step towards capturing etymological and word formation information between words in many languages. Supplementing the numerous lexical knowledge bases that focus on synchronic relationships, our resource aims at additionally capturing diachronic information by representing how words originated from other previously existing words. By navigating a network that captures both synchronic and diachronic relationships, as exemplified in Figure 1, one can easily see that the English "*doubtless*" is derived from "*doubt*", which in turn comes from Old French "*douter*", which evolved from the Latin word "*dubitare*". Additionally, starting from these latter nodes, further cognate forms are then easily discovered.

The information in the Etymological Wordnet is taken from Wiktionary, a well-known collaboratively edited online dictionary. While Wiktionary dumps are readily available, extracting a network of etymological information requires significant effort, as much of the etymological information is given in prose.

This work was supported in part by the National Basic Research Program of China Grants 2011CBA00300, 2011CBA00301, and NSFC Grants 61033001, 61361136003.

2. Background

In the 19th century, numerous connections between Indo-European languages were recognized, resulting in important insights that fundamentally shaped linguistics and anthropology. For instance, English "*ten*", German "*zehn*", Latin "*decem*", Greek "*deka*", and Sanskrit "*daśa*" are all cognates, i.e., words that descend from the same Proto-Indo-European ancestor. Due to various phonetic, phonological, and other changes, the word's pronunciation diverged in different communities, which came to have separate languages. Words may also evolve within what one typically would regard as stages of the same language, e.g., through sound changes such as the Great Vowel Shift in English, or more recently e.g. due to spelling reforms.

Language contact is another important factor. Languages may borrow words from one another, e.g. the English word "*cafe*" was borrowed from French "*café*". It is well-known that the English language has an unusually large number of words that were borrowed from Romance languages, often via Anglo-Norman, e.g. "*table*", "*bottle*", "*air*", "*choice*". The Etymological Wordnet data does not explicitly distinguish loanwords from etymological developments over time within a language or language family. However, with relevant background knowledge, e.g., the fact that Modern English developed from Middle English etc., one can recover this distinction to some extent.

Finally, when tracing the origins of words, synchronic word formation connections, in particular derivational and compositional links, are also important because many words come into existence via quite regular processes of affixation or compound formation. Note that such words may nevertheless enter the language at a particular point in time, as e.g. the case for the word "*website*". This point in time may be much later than the time the components that make up the new form entered the language. Also, note that such words may still have a non-compositional meaning

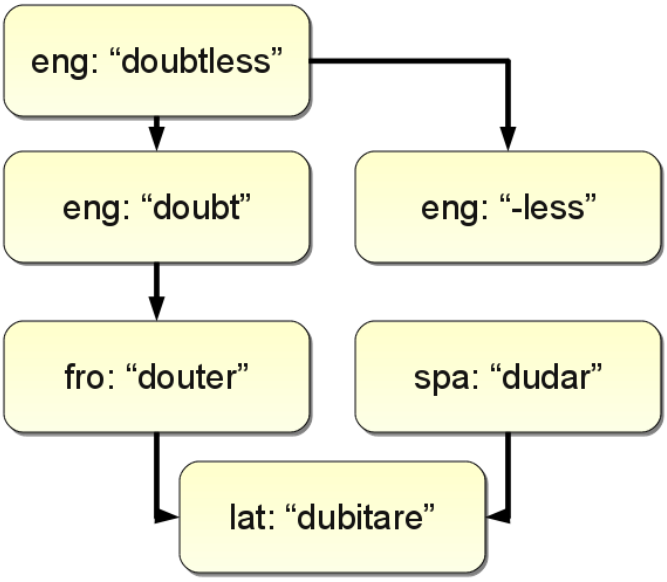

Figure 1: Excerpt from Etymological Wordnet

that cannot straightforwardly be inferred from the source morphemes. Examples of this include “*sexist*” (coined in the 1960s in analogy to “*racist*”) and “*microwave*” (the food-related meaning is only clear from the full form “*microwave oven*”).

3. Related Work

The study of etymology has a long history, and there are obviously numerous large etymological reference works that have appeared in print. For instance, for the English language, one might consult “The Concise Oxford Dictionary of English Etymology” (Hoad, 1993). Recently, some of these reference works, e.g. the ones in the Leiden Indo-European Etymological Dictionary series, have also been made available as databases. Unfortunately, other than the resources listed below, we are not aware of any open, freely available machine-readable versions of such works. Additionally, most such reference works are restricted to a single language or a set of closely related languages.

A notable exception is the Tower of Babel project by Sergei Anatolyevich Starostin, which provides a large and valuable database of etymological entries (cf. *starling.rinet.ru*). While machine-readable at a coarse-grained level, the data, however, is not represented as an easily navigable network of words as in the Etymological Wordnet. Additionally, some of the entries in the database are not generally accepted.

The World Loanword Database (WOLD) (Haspelmath and Tadmor, 2009) is another lexical resource that has been published as Linked Open Data and describes loanwords in 41 languages. For a set of 1,460 pre-selected meanings, the resource lists relevant words in these language and marks whether there is any evidence for borrowing from another language. If so, the donor language and word is given. Compared with the Etymological Wordnet, this project focuses on linguistic credibility by characterizing the amount of evidence for a borrowing and providing authorship information. The meaning-based structuring also means that this project better accounts for homonymy. However, despite its significant size, the WOLD does not aim at being a broad-scope resource. Unlike the Etymological Wordnet, it covers interesting minority languages like Saramaccan. However, it does not contain vocabularies for French or Spanish, for example. Its English vocabulary describes 1,505 words, while the Etymological Wordnet’s reliance on the English Wiktionary means that English and other major languages are covered to a significantly greater extent.

Numerous Swadesh lists (Swadesh et al., 1971) have been collected in machine-readable form. While these frequently list related forms side by side and can thus be useful for etymological research, the lists do not specifically mark whether two given words are cognates or not.

AfBo (Seifart, 2013) describes around 100 cases of affix borrowings between languages. For these, it contains extensive background information and references.

Finally, there are numerous lexical resources that describe morphological information within languages. While the Etymological Wordnet does cover salient derivational and compositional links, as a static database of relationships between forms, it cannot describe the full (often infinite) range of possibilities for word formation within a given language.

4. Approach

4.1. Model

The Etymological Wordnet attempts to describe word origins in terms of relationships between two terms, where the two terms may be in different languages. It is in this sense that the Etymological Wordnet is a network of words. Unlike the Princeton WordNet, it currently does not capture any word sense-specific information.

Information that they cannot directly capture faithfully can still be retained in textual form, e.g. using additional relationship attributes or meta-data. Fortunately, most forms of etymological information, including e.g. when a word’s use

English [edit]

Alternative forms [edit]

- dout (*obsolete*)

Pronunciation [edit]

- enPR: dout, IPA(key): /daʊt/
- Rhymes: -aʊt
- Audio (US) 0:00 MENU

Etymology 1 [edit]

From Middle English *douten*, from Anglo-Norman *douter*, from Old French *douter*, from Latin *dubitare*. Replaced Middle English *tweonien* ("to doubt") (from Old English *twēonian*, compare Old English *twēo* ("doubt, duplicity")). The modern spelling is probably under the influence of Middle French *doubter*.

Verb [edit]

doubt (*third-person singular simple present* **doubts**, *present participle* **doubting**, *simple past and past participle* **doubted**)

1. (*transitive, intransitive*) To lack confidence in; to disbelieve, question, or suspect. [quotations ▼]

   *He **doubted** that was really what you meant.*

Figure 2: Excerpt from Wiktionary article on "*doubt*", which explains the etymological roots going back to the Latin "*dubitare*"

was first attested, historic examples of a word's use, or even the presence of multiple conflicting etymological hypotheses could easily be couched in a machine-readable graph representation without resorting to textual comments.

4.2. Knowledge Extraction

The knowledge base is mined from the English version of Wiktionary using custom pattern matching techniques. We extract information from several different parts of Wiktionary.

Etymology Sections. We process the XML dump of Wiktionary, and segment articles by language-specific sections, since a single article can cover unrelated words in different languages. The "Etymology" subsections within them may contain arbitrary text describing the historical roots of a word, which means that they are not conveniently amenable to automated processing. Fortunately, certain general practices have become somewhat established. An example of this is given in Figure 2, where we see multiple parts starting with the word "*from*", followed by a language name and the actual word. Sometimes, etymology-specific templates are used to generate this code, which can facilitate automated processing even more. Our approach is to recursively parse the text using a set of regular expressions that cover many of the etymological patterns typically employed in Wiktionary. Such regular expressions extract the language (if mentioned), the original term, and the rest, i.e. the next element in an etymological chain.

Appendices. We also extract information from the Appendices of Wiktionary, which include pages for reconstructed words and roots in proto-languages like Proto-Indo-European. These include specific listings of etymological descendants. Parsing them requires interpreting the language names and list structures.

Gloss References. Sometimes, a word is not given its own genuine Etymology section, but just a quick reference in its gloss. The glosses often hold links to root forms for derivations, or links to standard forms when there are orthographic variations or other alternative forms. For instance, the English word "*booking*" is linked to the verb "*to book*".

Related Forms Sections. Many articles also have separate sections listing derived forms or alternative spellings, which we harvest as well.

Manual Additions. A small number (∼100) of manual additions have been made to the Etymological Wordnet.

4.3. Metadata

Due to space constraints, dictionaries appearing in print often refrain from providing references to the sources of their etymological information. As a computational resource, the Etymological Wordnet is not subject to such constraints and thus references the Wiktionary page that provided the information. This is particularly important because frequently, the source is not the page for the word itself, but rather some other page that references that word while tracing a longer etymological history. For example, the etymological link from Anglo-Norman "*estorie*" back to the Latin "*historia*" is found on the page for the English word "*story*".

Wiktionary pages in turn may reference the original sources of the etymological information they provide, though currently such citations are typically still lacking.

Another issue arising in etymology is that some words are unattested and only known as reconstructed forms. This information is captured as well.

Table 1: Coverage of the Etymological Wordnet

Relationship	Number of Entries
Etymological origin	523,758
Etymologically related	569,341
Derivational/compositional origin	2,342,027

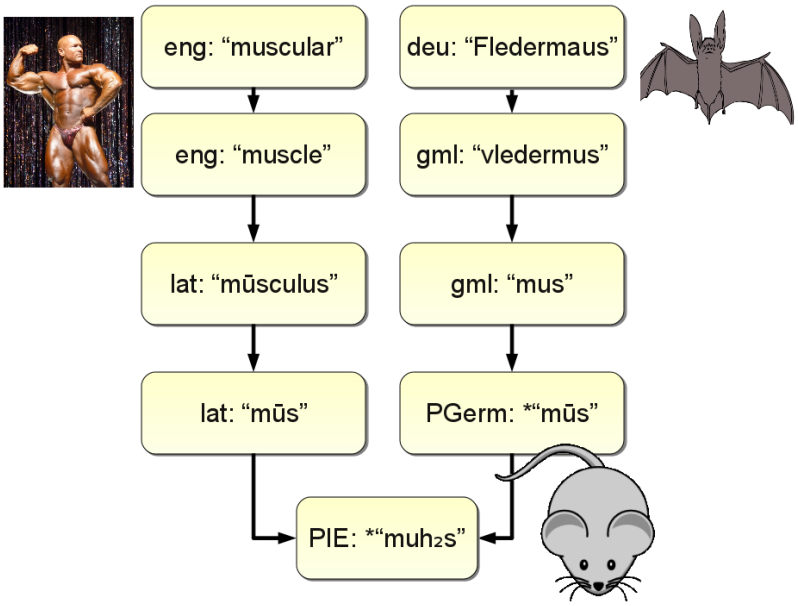

Figure 3: Connections in Etymological Wordnet

4.4. Cleaning

During the extraction phase, we parse the markup for internal links in order to obtain the actual word. We also need to support several special templates that are used on Wiktionary to embed links to words in various scripts and languages. Characters encoded using HTML entities are decoded as well. Terms are normalized by removing superfluous spaces.

Finally, we remove any duplicate entries, taking into account that the same word may have multiple Unicode encoding variants.

Additionally, we use a graph search algorithm to remove redundant links that are already indirectly provided by longer chains of links. The extractions come from different pages, which may vary in their levels of granularity. For instance, one page may trace a German word directly back to Old High German, while another may include an intermediate form in Middle High German. In such cases, we wish to remove the direct connection to Old High German if the Middle High German word already indirectly provides this connection.

5. Results

5.1. Statistics

We ran our extraction system on the 2013-09-07 version of the English Wiktionary. The resulting lexical network has over 3,000,000 terms. These terms are connected by 500,000 etymological origin links, 500,000 links for etymologically relatedness, and 2,300,000 derivational/compositional links between terms (see Table 1).

An etymological origin links connects a term to one or more source forms that gave rise to the term. Note that Wiktionary does not always make a clear distinction between synchronic word formation links (derivational or compositional ones) and genuine diachronic relationships. Etymology sections in Wiktionary may describe various forms of word origins, including derivational and compositional ones in some cases. The convention is that these sections “provide factual information about the way a word has entered the language”.[1] In this regard, our knowledge base simply follows Wiktionary’s policy and thus among the etymological origin links there are also significant numbers of synchronic word formation links. Note however that Wiktionary does aim at capturing the genuine historical origin of a word. Thus “*astrology*” is linked to its Ancient Greek ancestor, while the much more recent classical compound “*biology*” is connected to the affixes “*bio-*” and “*-logy*”.

In addition to etymological origin links, our data also contains etymological relatedness links. Etymological relatedness can be regarded as a generalization that includes etymological origin links but also connections between cognate forms.

While there are a small number of incorrectly decoded words, overall the precision of the resource is roughly 100% with respect to Wiktionary as the ground truth. While

[1] Source: `http://en.wiktionary.org/wiki/Wiktionary:Etymology` (as of 2014-03)

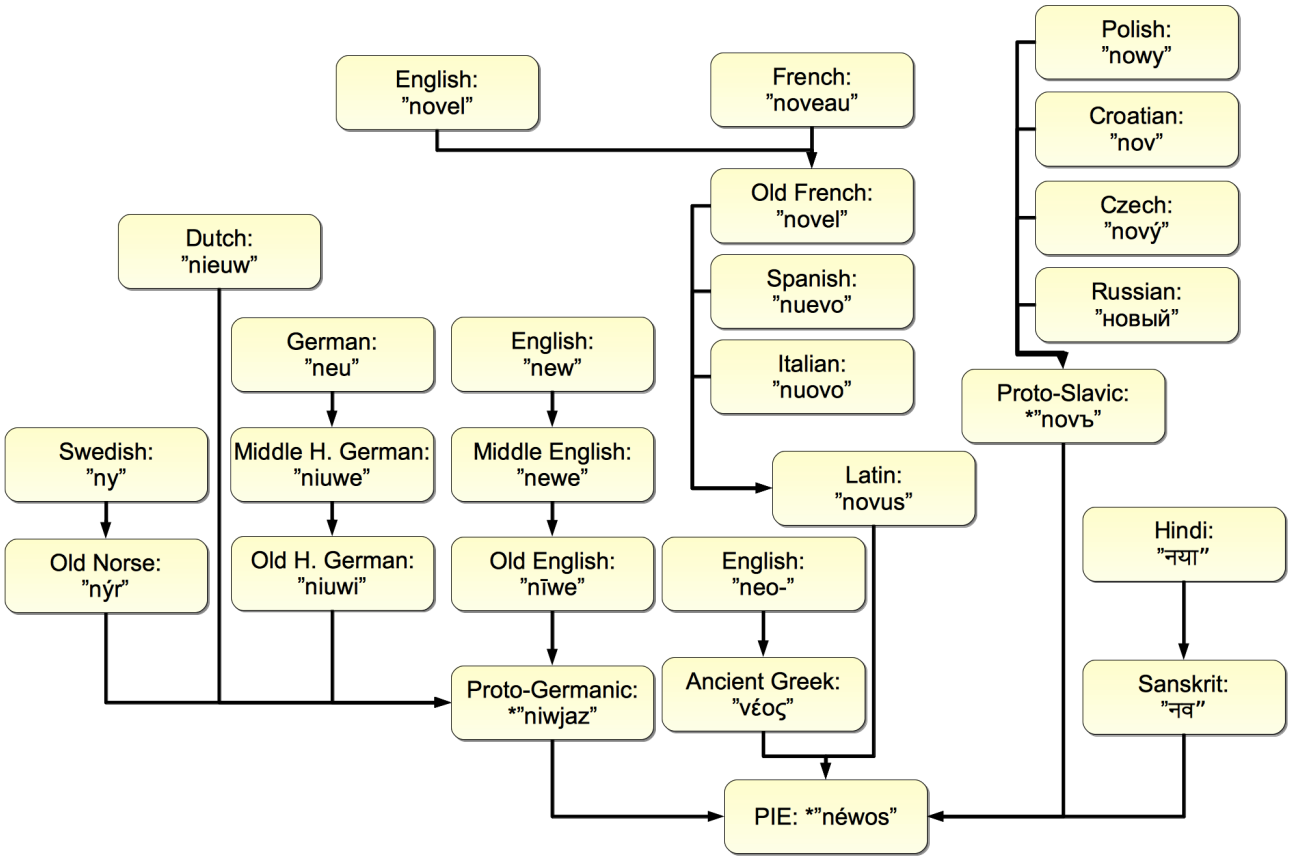

Figure 4: Descendants of a Word in Etymological Wordnet

Wiktionary of course allows contributions from the general public, we hypothesize that etymological entries are typically entered by users with at least some basic familiarity with etymology. Still, there is a risk is that such contributors may present false hypotheses or even folk etymologies as uncontested truths. Wiktionary and in extension the Etymological Wordnet thus do not necessarily constitute credible sources for scholarly research on individual etymologies. However, as long as this fact is kept in mind, they can be used as exploratory tools and for computing general macro-level tendencies.

Within this data, one can for instance discover relationships like the ones in Figure 3, where the connection between the English word "*muscle*" and the German word for bats ("*Fledermaus*") is revealed. Once discovered, one can then verify such connections using more authoritative sources if necessary. Figure 4 shows another excerpt, in which a sample of some of the descendants of the Proto-Indo-European reconstruction " *néwos*" are displayed.

The Etymological Wordnet can also be queried in conjunction with UWN (de Melo and Weikum, 2009), which has been extended to incorporate language family data extracted from Wikipedia and other sources into the hypernym hierarchy of Princeton WordNet (de Melo and Weikum, 2010). Figure 5 illustrates a query that aims at finding words in West Germanic languages with origins in the Austronesian language family. An example would be the English word "*orangutan*", which has its roots in Malay. The resulting data can also be used for statistical analyses. For instance, Table 2 lists the most common (immediate) source languages for a small selection of languages.

5.2. Data Access

We have created an RDF version of this data, relying on the term URIs defined by the Lexvo.org service (de Melo and Weikum, 2008; de Melo, 2014).

Existing standards like TEI P5 (Burnard and Bauman, 2009) define a semi-structured representation of etymological data, rather than a genuinely structural one that exposes relationships between words using a network-like graph model. Graph representations expose the connections between words much more explicitly. Due to affixes such as "*non-*", "*-ize*", etc., it turns out that much of the graph actually constitutes a single connected component that can be navigated by following links. In addition, graph representations are machine-readable and more language-neutral, which makes them reusable in different contexts.

We provide a Java library (de Melo and Weikum, 2012) that makes it easier to query the data in natural language processing tools. In fact, an earlier version of the Etymological Wordnet, with significantly less data, has already been successfully used for cross-lingual text classification (Nastase and Strapparava, 2013).

5.3. Discussion

The Etymological Wordnet is an important project that we believe can be useful for Digital Humanities research. It has also already proven useful in NLP tasks, although this

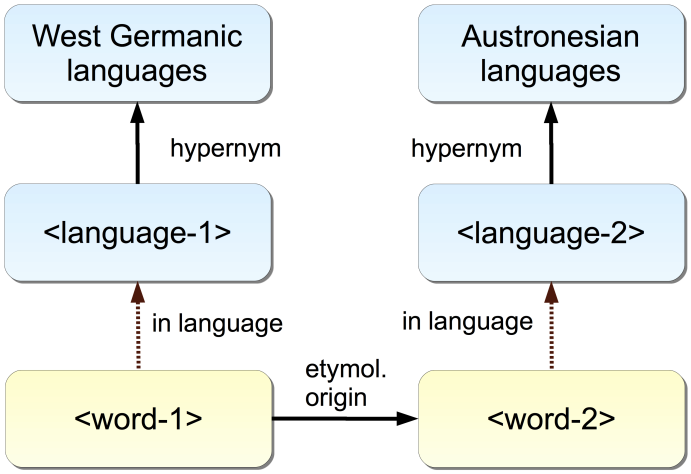

Figure 5: Queries using Language Family Information

Table 2: Top Etymological Source Languages

Language	Source Languages
English	1. Latin 2. Middle English 3. French 4. Old French 5. Ancient Greek
German	1. Old High German 2. French 3. Middle High German
French	1. Latin 2. Old French 3. Middle French
Italian	1. Latin 2. French 3. Ancient Greek
Spanish	1. Latin 2. French 3. Ancient Greek
Icelandic	1. Old Norse 2. English 3. Danish
Irish Gaelic	1. Old Irish 2. English 3. Middle Irish

was not a goal of the project.

However, given the ambitious scope of this project, its coverage still remains quite low in comparison to the large etymological reference works that have appeared in print for specific languages and language families. We expect that the coverage of this resource will continue to grow as Wiktionary gets updated and possibly other sources are added. At the same time, there is also a risk that the growth of Wiktionary's Etymology sections will entail the use of language that is harder to parse automatically.

Another long-term desideratum would be allowing for manual additions of semantic descriptors. This would enable the resource to describe semantic change in etymological relationships. Additionally, semantic change of a word within a language could be described as well (Sweetser, 1990). Such additions would also allow the Etymological WordNet to become more like lexical semantic wordnets such as the Princeton WordNet.

6. Conclusion

We have presented the first broad-coverage etymological database that aims at making word relationships across a large number of human languages available in machine-readable form. We are currently in the process of extending the coverage of the resource by extracting from a greater range of linguistic patterns. While much remains to be done in this area, the Etymological Wordnet has already proven useful for natural language processing.

Acknowledgements

This work was supported in part by National Basic Research Program of China Grants 2011CBA00300, 2011CBA00301 and National Natural Science Foundation of China Grants 61033001, 61361136003.

7. References

Burnard, L. and Bauman, S., (2009). *TEI P5: Guidelines for Electronic Text Encoding and Interchange, Version 1.4.1*. TEI Consortium.

de Melo, G. and Weikum, G. (2008). Language as a foundation of the Semantic Web. In Bizer, C. and Joshi, A., editors, *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC 2008)*, volume 401 of *CEUR WS*, Karlsruhe, Germany. CEUR.

de Melo, G. and Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. In *Proc. 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*. ACM.

de Melo, G. and Weikum, G. (2010). Towards universal multilingual knowledge bases. In Bhattacharyya,

P., Fellbaum, C., and Vossen, P., editors, *Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the 5th Global WordNet Conference (GWC 2010)*, pages 149–156, New Delhi, India. Narosa Publishing.

de Melo, G. and Weikum, G. (2012). UWN: A large multilingual lexical knowledge base. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 151–156, Stroudsburg, PA, USA. Association for Computational Linguistics.

de Melo, G. (2014). Lexvo.org: Language-related information for the linguistic linked data cloud. *Semantic Web Journal*.

Haspelmath, M. and Tadmor, U., editors. (2009). *WOLD*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Hoad, T. F. (1993). *The Concise Oxford dictionary of English etymology*. Oxford University Press Oxford.

Nastase, V. and Strapparava, C. (2013). Bridging languages through etymology: The case of cross language text categorization. In *Proceedings of ACL 2013*.

Seifart, F., editor. (2013). *AfBo: A world-wide survey of affix borrowing*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Swadesh, M., Sherzer, J., and Hymes, D. (1971). *The Origin and Diversification of Language*. Routledge and K. Paul.

Sweetser, E. (1990). *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Number v. 54 in Cambridge Studies in Linguistics. Cambridge University Press.