etytree

A graphical and interactive etymology dictionary based on Wiktionary

Extended Abstract†

Ester Pantaleo
Wikimedia Foundation grantee
esterpantaleo@gmail.com

Vito Walter Anelli
Dipartimento di Ingegneria Elettrica e dell'Informazione
Politecnico di Bari
Italy
vitowalter.anelli@poliba.it

Tommaso Di Noia
Dipartimento di Ingegneria Elettrica e dell'Informazione
Politecnico di Bari
Italy
tommaso.dinoia@uniba.it

ABSTRACT

Etymological definitions in the English version of Wiktionary are particularly well compiled and contain a very rich set of information hindering binary, i.e., etymological, relationships between words, e.g., “word A derives from word B, word B derives from word C, etc”.

The etytree project (http://tools.wmflabs.org/etytree/) extracts this information into a database of triples or data entities composed of subject-predicate-object (where a possible predicate is “derives from”). The project uses this database to create an intuitive and multilingual graphical etymology dictionary.

Potentially, the tool can be integrated into Wikidata when the Wikidata-for-Wiktionary proposal turns into production.

KEYWORDS

etymology, Wiktionary, natural language processing

1 INTRODUCTION

The project etytree consists of a tool to extract etymological relationships between lexemes contained in the Etymology section of the English Wiktionary and of a graphical tool to explore the extracted database of etymological relationships.

1.2 THE DATABASE

The project is based on DBnary[1], a tool that extracts Definition, Part of Speech, Synonyms, etc from Wiktionary pages. The extraction tool is available at https://bitbucket.org/esterpantaleo/dbnary_etymology and uses regular expressions and parsing of Wiktionary templates.

The data has been collected into a RDF database of resources/nodes (words) and properties/links (etymological relationships) that is synchronized with the English Wiktionary and can be queried at http://etytree-virtuoso.wmflabs.org/sparql.

1.3 THE GRAPHICAL INTERFACE

The graphical interface to the database (or graphical etymology dictionary) can be explored at http://tools.wmflabs.org/etytree/etymology/resources/html/index.html and represents the extracted etymological relationships as well as the associated lexical information using graphs. It uses d3.js (https://d3js.org/), a JavaScript library for manipulating documents based on data and infers the tree structure from the RDF database on the fly through specific SPARQL queries using Virtuoso (https://virtuoso.openlinksw.com/) as a Database Management System.

1.3 SIMILAR WORK

A similar etymology extraction tool is the Etymological Wordnet (www1.icsi.berkeley.edu/~demelo/etymwn/) [2], [3]. Unfortunately the Etymological Wordnet is not publicly available and, from a first inspection of data extracted using it (which is available at the link above), we believe etytree can extract more etymological relationships.

2 IMPACT OF THIS WORK

The interest of this tool lies in its potential for Wiktionary users, editors and for researchers or more generally people interested in languages and etymologies.

2.1 IMPACT ON USERS

With this tool, users can discover new words when they search for a specific etymological definition, e.g., they can discover words that derive from the same ancestral word, both in their own language and in other languages. This happens in an intuitive way without having to read fairly long and complex sentences that describe etymological relationships between words and without having to navigate across multiple Wiktionary pages.

2.2 IMPACT ON EDITORS

Editors can easily spot inconsistencies between etymological relations described across multiple Wiktionary pages using the visualization of the whole etymological tree.

2.3 IMPACT ON RESEARCHERS

Researchers can use the database of etymological relationships to study etymologies on a large scale. Potentially, they could extend the database of etymological relationships to include semantics or pronunciations, to study how they evolve through time across etymological trees and across languages.

3 CONCLUSIONS

Because of its nature, we believe this work will attract new users to Wiktionary and will improve as well as increase its content. It will also encourage Wiktionary editors to use clear patterns and standard rules to format etymologies, e.g. conflicting etymologies (but also Wiktionary sections in general). We will encourage discussions on wiki pages and on the project wiki page. This process will hopefully help to turn Wiktionary into a machine readable resource and therefore into a database without an important loss of information.

The database produced by this project is ready to be exported to Wikidata when the Wikidata-for-Wiktionary proposal (see https://www.wikidata.org/wiki/Wikidata:Wiktionary/Development/Proposals/2015-05) turns into production.

Because of the complexity of the data (Etymology Sections) the extracted database contains some incorrect entries. We hope that users will contribute to Wiktionary to spot those inconsistencies. We would like to work together with them to improve etymology sections in Wiktionary and etytree simultaneously.

3.1 FUTURE DEVELOPMENTS

As the structure of etymological trees (or graphs) is language independent, this project could be extended to use more language versions of Wiktionary, although etymologies seem rather incomplete/informal in other languages. Furthermore, the textual part of the tree (definition of words, language tags, etc) could be translated into different languages which would make this tool international and language independent, thus considerably extending its scope.

ACKNOWLEDGMENTS

This work is supported by the Wikimedia Foundation through an IEG grant (https://meta.wikimedia.org/wiki/Grants:IEG/A_graphical_and_interactive_etymology_dictionary_based_on_Wiktionary) to Ester Pantaleo.

REFERENCES

[1] Sérasset Gilles (2014). DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. to appear in Semantic Web Journal (special issue on Multilingual Linked Open Data).

[2] Gerard de Melo, Etymological Wordnet: Tracing the History of Words,In: Proc. LREC 2014. ELRA, 2014, Paris, France.

[3] Gerard de Melo and Gerhard Weikum, Towards Universal Multilingual Knowledge Bases, In: Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the 5th Global Wordnet Conference (GWC 2010), Narosa Publishing 2010, New Delhi India.