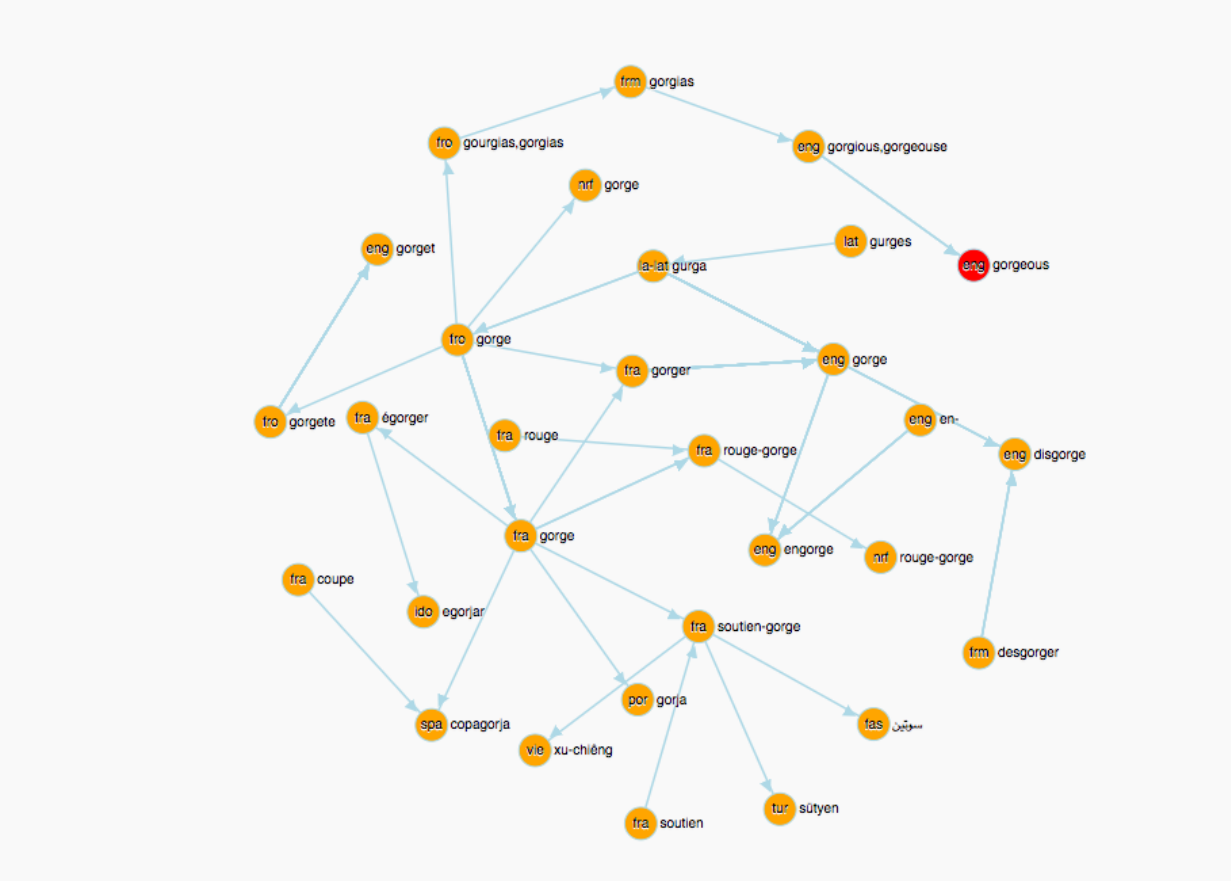

Figure 1: A screenshot of the interactive visualization produced by `etytree` for the English word "*gorgeous*". From the graph it is possible to see that the English words "*gorgeous*", "*disgorge*", "*gorget*", and archaic words "*gorge*" and "*engorge*" are etymologically related.

A project similar to `etytree` is Etymological Wordnet[4] [2, 3] which is, unfortunately, neither publicly available nor maintained anymore.

2. THE MODEL

The `etytree` extraction tool uses regular expressions and parsing of both `Wiktionary` templates and links. It assumes a standard structure for the different sections containing etymologies, i.e., the Etymology section, the Derived terms section, the Descendants section, the namespace with Reconstructed Terms (still in the works), the `etymtree` template[5].

2.1 Etymology sections

Figure 2 presents a screenshot of the Etymology section of English word "*gorgeous*" in `English Wiktionary`. The same section in the `xml` dump (our data source) as well as in the edit tab of the online `English Wiktionary` is:

===Etymology===
From Early Modern English {{m|en|gorgious}}, {{m|en| gorgeouse}}, from {{etyl|frm|en}} {{m|frm|gorgias||elegant, fashionable}}, from {{etyl|fro|en}} {{m|fro|gourgias}}, {{m|fro|gorgias|| gorgeous, gaudy, flaunting, gallant, fine}}, of uncertain formation, but apparently connected with {{cog|fro|gorgias||a gorget, ruffle for the neck}}, from {{etyl|fro|en}} {{m|fro|gorge||bosom, throat}}. See {{l|en|gorge}}. Sense evolution was probably that of "swelling of the throat or bosom due to pride, bridling up" to "assume an air of importance, flaunting".

After inspection of many different Etymology sections we inferred a set of recurrent patterns that we constructed using regular expressions. The most common pattern is[6]:

(FROM)?(LANGUAGE LEMMA |LEMMA)(COMMA |DOT |OR)

Using this pattern plus a set of rules we extract etymological relationships into a `RDF` database. In what follows we present some examples of rules that we use.

If we find a match to the pattern above with DOT or OR in the last group, we ignore all the text following the match. We ignore anything after a dot (DOT) because generally Etymology sections start with a chain of etymological relationships followed by a dot and then contain some descriptive text that is not easily parsable. We ignore anything following OR (alternative etymologies) as alternative etymologies are not presented in a standard format in the `English Wiktionary`. We also ignore anything that follows a match to SUPERSEDED BY[7] or COGNATE TO[8].
Also we use a pattern to match compounds, i.e., sentences like

{{m|en|door}}+{{m|en|bell}}

or

Compound of {{m|en|door}} and {{m|en|bell}}

Whenever we find a match to a compound pattern, we ignore everything after the match, as there is no standard for the etymology of compound words.

While the selected patterns generally correctly reflect real patterns (as Etymology sections use very well defined standards[9]), some etymologies are written in non-standard ways, which implies that the corresponding extraction is incorrect (or partially incorrect). We are trying to interact with the community of editors of `English Wiktionary` to better understand the standards they use and to encourage the use of more standards that would allow the community to have a lower amount of data loss and a lower rate of incorrectly extracted etymological relationships.

One example of non-standard Etymology sections uses links instead of templates to represent words that are etymologically related (e.g. [[door]] instead of {{m|en|door}}). This is a major problem because in Etymology sections words with links often correspond to descriptive words or glossary, for example the Etymology section of "*Davidsen*" is:

===Etymology===
Originally a [[patronymic]] from {{suffix|David|sen|lang=da}}.

and clearly "*patronymic*" here is not etymologically related to "*Davidson*". In this particular case, a standard that encourages the use of links to the glossary for words like "*patronymic*", i.e. [[Appendix:Glossary#patronymic|patronymic]], (and for "*ablative*", "*zero-grade*", etc.) in Etymology sections would help automatic data extraction.

Other lexemes that usually have non-standard Etymology sections are phrases. For example "*until the cows come home*" has the following Etymology section:

===Etymology===
Possibly from the fact that [[cattle]] let out to pasture may be only expected to return for milking the next morning; thus, for example, a party that goes on " until the cows come home" is a very long one. Alternatively, the phrase may have a Scottish origin,<ref>See, for example, {{cite-web|title=Till the cows come home |url=http://www.phrases.org.uk/meanings/382900.html |archiveurl=https://web.archive.org/web/20160611134612/ http://www.phrases.org.uk/meanings/382900.html | archivedate=11 June 2016 |work=Phrase Finder | accessdate=30 March 2013}}.</ref> and may derive from the fact that cattle in the [[w:Scottish Highlands|Highlands]] are put out to graze on the [[common#Noun|common]] where grass is plentiful. They

[4] `www1.icsi.berkeley.edu/~demelo/etymwn/`
[5] See `https://en.wiktionary.org/wiki/Template:etymtree`
[6] where FROM can be any of the following:

"[Ff]rom", "[Bb]ack-formation (?:from)?", "[Aa]bbreviat(?:ion|ed)? (?:of|from)?",

and many more, LANGUAGE corresponds to the `etyl` template, LEMMA corresponds to different templates in practice (e.g. `m`, `l`, etc, generally embedding lexemes) or wiki links, COMMA corresponds to ",", DOT corresponds to "." or ";", and OR corresponds to "or" (neither followed nor preceded by a character).

[7] Namely "[Ss]uperseded", v[Dd]isplaced(?: native)?", "[Rr]eplaced", "[Mm]ode(?:l)?led on", and more.
[8] Namely "[Rr]elated(?: also)? to", "[Cc]ognate(?:s)? (?:include |with |to |including)?", "'Ss]ee(?:n)? (?:also)?"), and more.
[9] See `https://en.wiktionary.org/wiki/Wiktionary:Etymology`

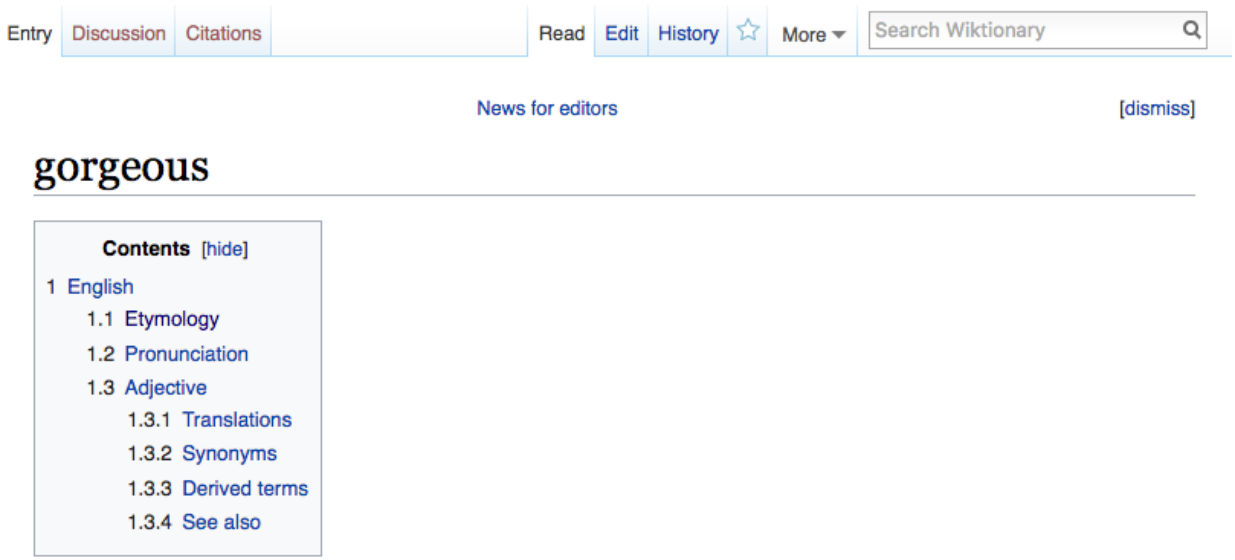

Figure 2: A screenshot of the English entry "*gorgeous*" in the `English Wiktionary`

stay out for months before scarcity of food causes them to find their way home in the autumn for feeding.

we propose to add template (for example {{detailed etymology}}) before long descriptive etymologies that don't have a standard chain of etymological relationships which would signal to the extraction algorithm to ignore that section.

In the current version, `etytree` parses links like [[cattle]] in the Etymology section above as an ancestor of "*until the cows come home*" and therefore infers an incorrect etymological relationship. We decided to keep those links for now, as we hope that editors will fix those entries and set a clear standard in the structure of Etymology sections.

2.2 Derived terms sections

Derived terms sections are pretty standard with some exceptions. Below we copy the Derived terms section of English "*gorgeous*", which is representative of how Derived terms sections are usually structured:

```
====Derived terms====
* {{l|en|gorgeously}}
* {{l|en|gorgeousness}}
```

2.3 Descendants sections

Descendants sections also are written in a standard way (with some exceptions). Below we copy the (beginning of the) Descendants section of Latin "*aqua*":

```
====Descendants====
* Eastern:
** Aromanian: l|rup|apã
** Istro-Romanian: l|ruo|âpe
** Megleno-Romanian: l|ruq|apu
** Romanian: l|ro|apă
* Franco-Provençal: l|frp|àiva
* Gallo-Italian:
** Emilian: l|egl|âcua
** Ligurian: l|lij|aigua, l|lij|ægoa
** Lombard: l|lmo|acqua, l|lmo|ègua
** Piedmontese: l|pms|eva
** Romagnol: l|rgn|aqua, l|rgn|acva
** Venetian: l|vec|aqua
```

2.4 Appendix with reconstructed words

Reconstructed terms are words, roots, or phrases that are not attested but have been reconstructed by linguists and are conventionally identified with an initial asterisk. They are defined in the namespace Reconstruction (see for example entry "$h_2k^weh_2$" defined at `https://en.wiktionary.org/wiki/Reconstruction:Proto-Indo-European/h%E2%82%82%C3%A9k%CA%B7eh%E2%82%82`) and are structured similarly to regular `Wiktionary` entries.

2.5 `etymtree` template

The `etymtree` template is a template used in `Wiktionary` to describe etymological trees and reflects the structure of the Descendants sections.

3. THE DATABASE

The database is installed on the `Wikimedia Labs`, the `Wikimedia Foundation`'s cloud computing environment. It is managed by `Virtuoso version 07.20.3217` on `Linux`.

The extracted database consists of 6 million distinct entries (6023380) in 3365 languages, with Latin having the highest number of entries (around 13% or 806999 entries), followed by English (9% or 547506), Italian (8.5% or 515059), Spanish (7% or 419889), Russian (5.5% or 331798), French (5% or 305973), Portuguese (4% or 244784), German (3% or 185520).

With appropriate queries to the `SPARQL` endpoint we can ask interesting questions.

For example, we can ask which languages English words derive from. The extracted English words derive mostly from other English words (2702), Middle English words (1152), Latin words (1116), French words (832), Old French words (715). Italian words derive mostly from Latin words (1132), Italian words (457), Spanish words (358), French words (147), Greek words (118). French words derive mostly from Latin words (2190), Middle French (1185), Old French (995), French (982), Italian words (584).

We can also ask `Virtuoso` to list the most connected entries. The most connected entries are affixes, namely English "*-ly*" (7070 connections), "*non-*" (6900 connections), "*un-*" (6873), "*-ness*" (5312). The most connected French affix is "*-ment*" (2573). Hungarian "*-ok-*" (2054), "*-ek-*" (1809),"*-k-*" (1821) and Italian "*-mente*" (2035), "*-ità*" (1670) are the most connected affixes in their respective languages.

The most connected entries that are not affixes are English lemmas "*man*" (353 connections), "*back*" (303), "*head*" (290), followed by "*work*", "*house*", "*wood*", "*land*", "*line*". These highly connected nodes slow down queries launched by the visualization tool. We are currently working on the design of more efficient queries given the available data.

4. CONCLUSIONS

We have presented `etytree`, a tool to visualize etymological relationships between words in the form of a connected graph. The tool is currently under development but a first working release is available at `http://tools.wmflabs.org/etytree/etymology/resources/html/index.html`.

This tool can be a valuable resource for people that are interested in the history of words or words in general as they can discover new words in other languages that are etymologically related to the searched words, as well as for etymology enthusiasts, as they can explore etymological relationships in a completely new way. Also we believe that the database can be a valuable resource for linguists as they can study etymologies on a much larger scale.

Because of its nature, we believe this work will attract new users to `Wiktionary` and will improve as well as increase its content. We also believe it will encourage `Wiktionary` editors to use more standard rules to format etymologies. We hope that this project will help to turn the whole `Wiktionary` into a machine readable resource with the minimum possible loss of information.

Because of the complexity of the original data contained in `Wiktionary` (especially the complexity of 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 even more `Wiktionary` Etymology sections and to improve `etytree` simultaneously.

The project has the potential to grow in both content and quality as it is open source and relies on data coming from a collaborative and multilingual resource as Wiktionary.

5. FUTURE DEVELOPMENTS

As the structure of etymological trees (or graphs) is language independent, this project could be extended to use etymological relationships described in other language versions of Wiktionary, although Etymology sections seem rather incomplete/informal in other languages (Russian might be the next target language).

In addition, as the textual part of the tree (definition of words, language tags, etc.) can be exported from different language versions of Wiktionary, this tool can easily become available in different languages, thus considerably extending its scope.

Last, this tool can be integrated into Wikidata when the Wikidata-for-Wiktionary[10] proposal turns into production.

6. ACKNOWLEDGMENTS

This work is supported by the Wikimedia Foundation through an IEG grant[11] to Ester Pantaleo. We would like to thank the Wiktionary and the Wikidata communities for their help and for their precious work.

7. AUTHOR CONTRIBUTIONS

Ester Pantaleo conceived the idea and developed the tool, Tommaso Di Noia contributed to the research project with monthly meetings, Gilles Sérasset helped integration with DBnary and provided some computational resources, Vito Walter Anelli helped formulating appropriate queries to the Virtuoso DBMS. Everyone participated in the writing of this paper.

References

[1] S. Gilles. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web Journal - Special Issue on Multilingual Linked Open Data*, 6(4): 355–361, 2015.

[2] G. de Melo. Proc. lrec. ELRA, Paris, France, 2014.

[3] G. de Melo and G. Weikum. Proceedings of the 5th global wordnet conference (gwc 2010). Narosa Publishing, New Delhi India, 2010.

[10] https://www.wikidata.org/wiki/Wikidata:Wiktionary/Development/Proposals/2015-05

[11] https://meta.wikimedia.org/wiki/Grants:IEG/A_graphical_and_interactive_etymology_dictionary_based_on_Wiktionary