Wiktionary
The free dictionary

etytree tool

database of lexical info + etymological relationships + interactive tool
based on the English Wiktionary

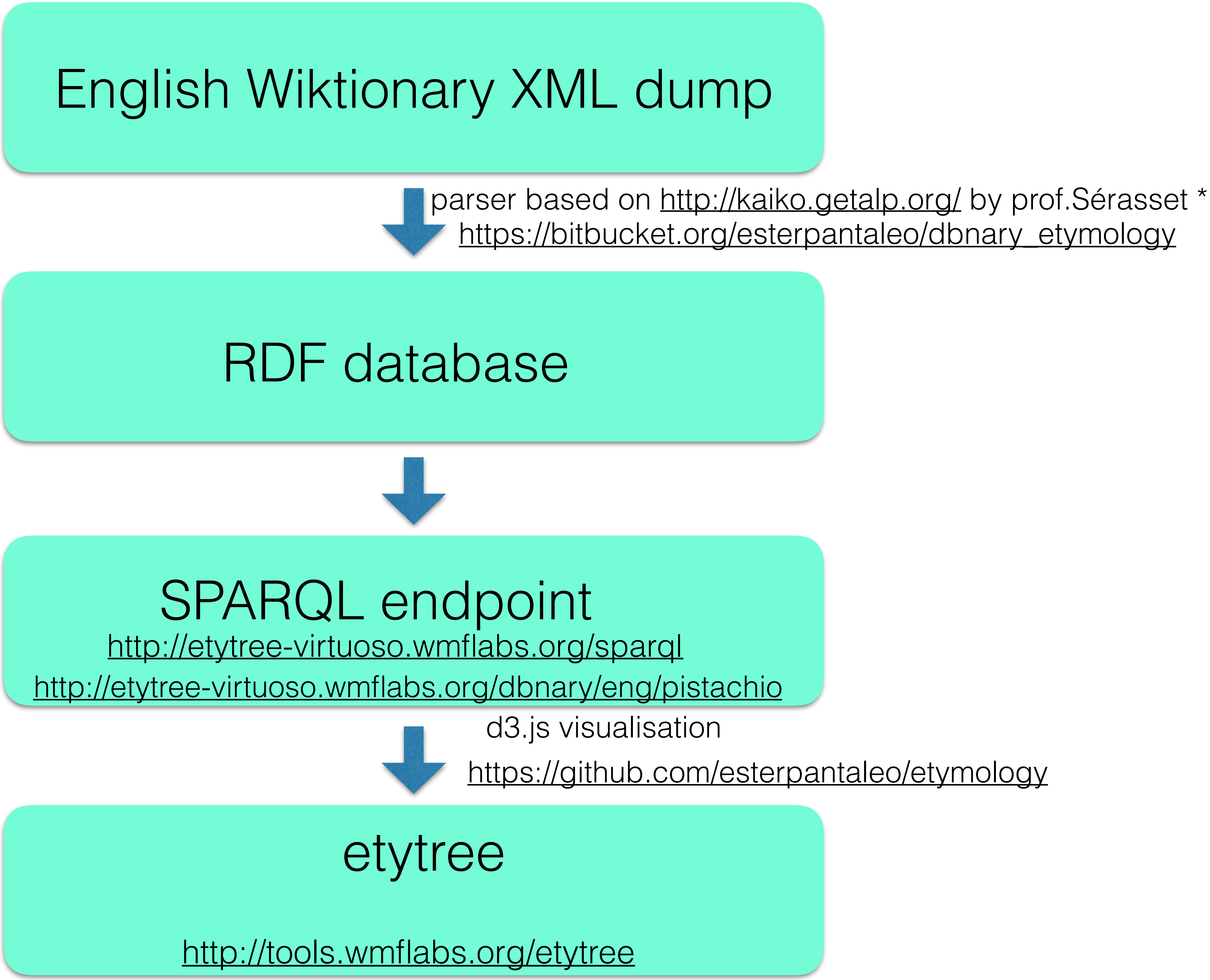
Epantaleo
WWW.EPANTALEO.COM
ESTERPANTALEO@GMAIL.COM

LIG
Laboratoire d'Informatique de Grenoble

de' remi facemmo ali
1990
POLITECNICO DI BARI

WIKIMEDIA
FOUNDATION

the workflow



English Wiktionary XML dump

parser based on http://kaiko.getalp.org/ by prof.Sérasset *
https://bitbucket.org/esterpantaleo/dbnary_etymology

RDF database

SPARQL endpoint
http://etytree-virtuoso.wmflabs.org/sparql
http://etytree-virtuoso.wmflabs.org/dbnary/eng/pistachio

d3.js visualisation
https://github.com/esterpantaleo/etymology

etytree

http://tools.wmflabs.org/etytree

* G. Sérasset *Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf*, Semantic Web, 2015

why the English Wiktionary?

- multilingual
- etymology sections in the English Wiktionary are relatively easy to parse with a machine:

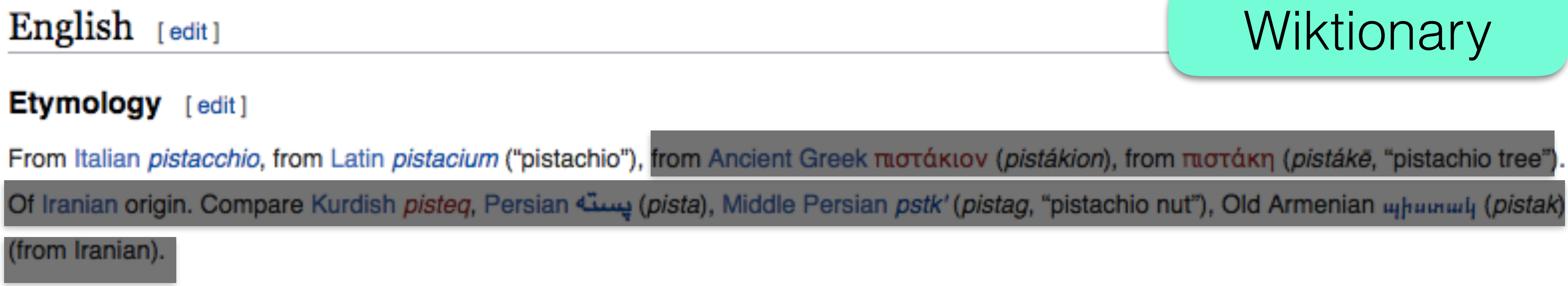
  +: follow a clear set of standard rules

  https://en.wiktionary.org/wiki/Wiktionary:Etymology

  -: there are many exceptions

  e.g.: use of links: ‘’[[fortia]]'' in Galician “forza”

Wiktionary

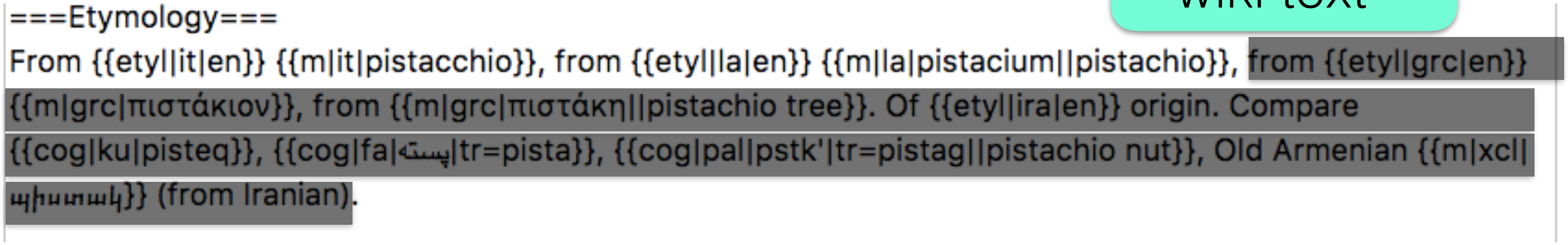


English [edit]

Etymology [edit]

From Italian *pistacchio*, from Latin *pistacium* ("pistachio"), from Ancient Greek πιστάκιον (*pistákion*), from πιστάκη (*pistákē*, "pistachio tree"). Of Iranian origin. Compare Kurdish *pisteq*, Persian پسته (*pista*), Middle Persian *pstk'* (*pistag*, "pistachio nut"), Old Armenian պիստակ (*pistak*) (from Iranian).

wiki text



```
===Etymology===
From {{etyl|it|en}} {{m|it|pistacchio}}, from {{etyl|la|en}} {{m|la|pistacium||pistachio}}, from {{etyl|grc|en}}
{{m|grc|πιστάκιον}}, from {{m|grc|πιστάκη||pistachio tree}}. Of {{etyl|ira|en}} origin. Compare
{{cog|ku|pisteq}}, {{cog|fa|پسته|tr=pista}}, {{cog|pal|pstk'|tr=pistag||pistachio nut}}, Old Armenian {{m|xcl|
պիստակ}} (from Iranian).
```

regex and grammar

FROM LANGUAGE LEMMA1 COMMA FROM LANGUAGE LEMMA2 COMMA […] DOT

from which the triples:

LEMMA etymologicallyDerivesFrom LEMMA1
LEMMA1 etymologicallyDerivesFrom LEMMA2
[…]

after substitution

eng:pistachio etymologicallyDerivesFrom ita:pistacchio .
ita:pistacchio etymologicallyDerivesFrom lat:pistachium .
[…]

RDF

E. Pantaleo, V.W. Anelli, T. Di Noia, G. Sérasset *Etytree: A Graphical and Interactive Etymology Dictionary Based on Wiktionary,* WWW'17 Companion, April 3–7, 2017, Perth, Australia. wikiworkshop.org/2017/papers/p1635-pantaleo.pdf

some statistics

Lexical entry (lemon ontolex)

“a lexical entry is a word, multiword expression or affix with a single part-of-speech, morphological pattern, etymology and set of senses”*

- 5,630,308 entries
- from 5,379,386 English Wiktionary pages
- in 3281 languages

Etymology entry

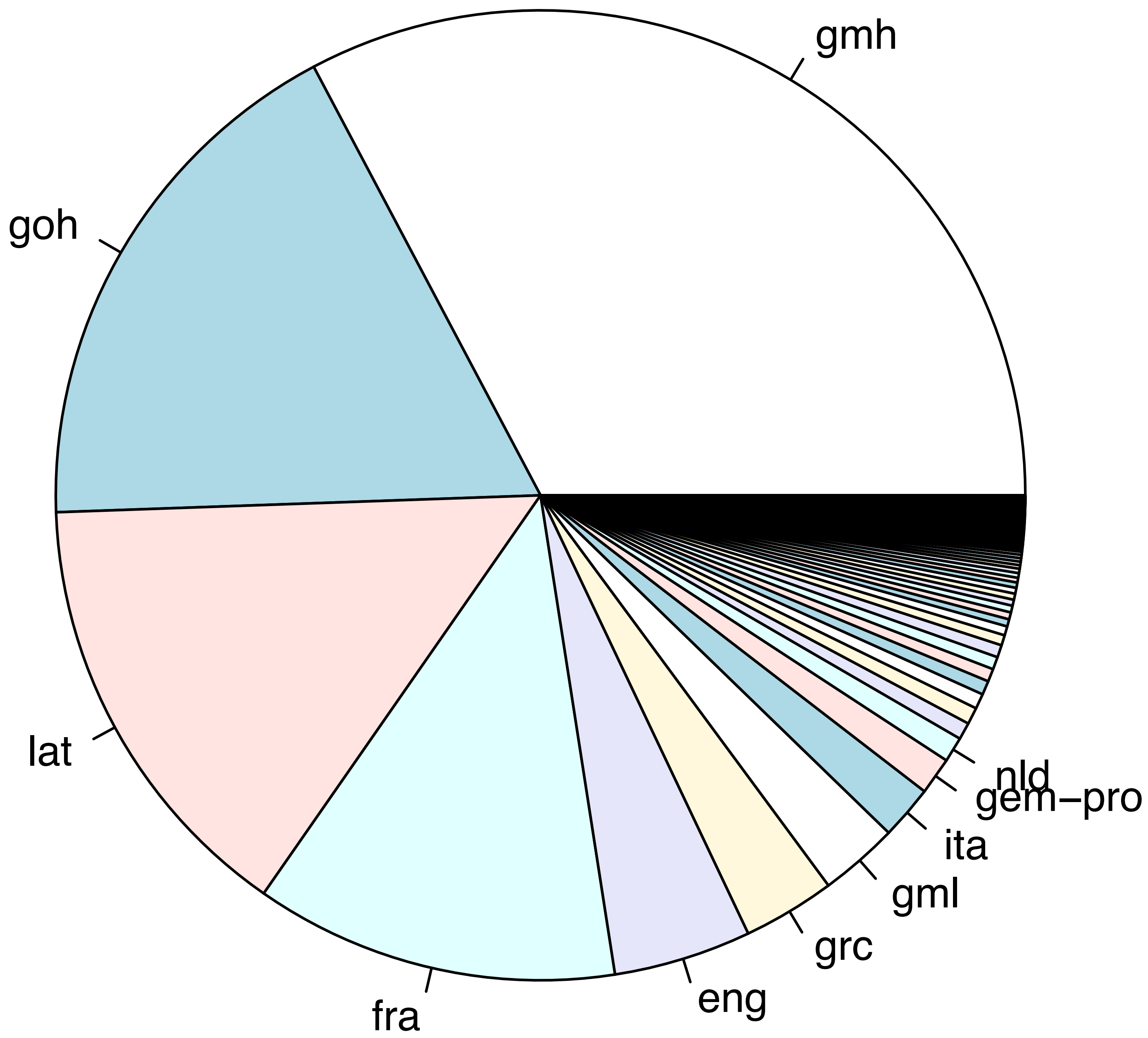
“a set of lexical entries that share the same etymology”

- 6,365,445 entries
- from 925,993 English Wiktionary etymology sections
- in 3973 languages

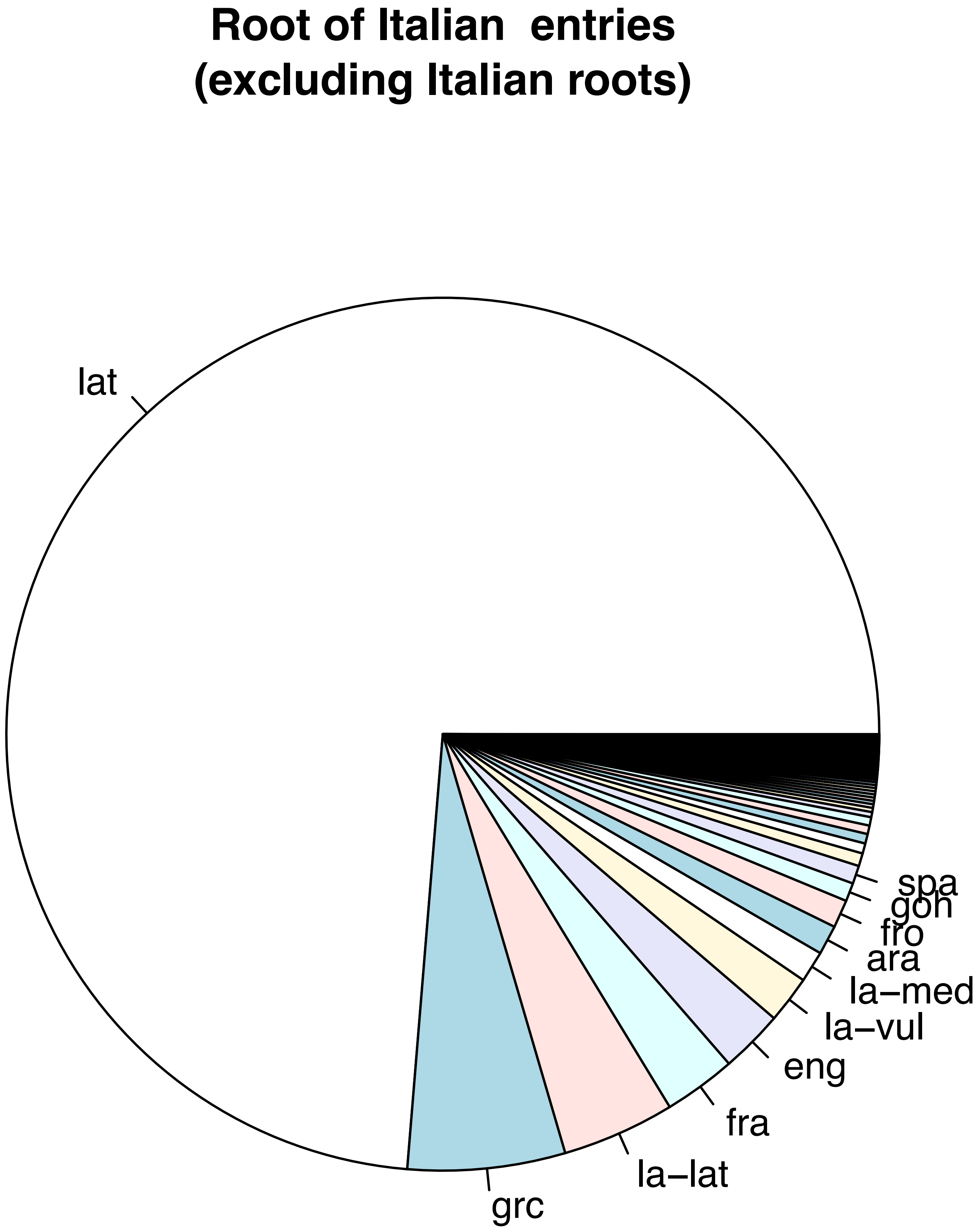
* https://www.w3.org/community/ontolex/wiki/Final_Model_Specification#Lexical_Entries

some statistics

Root of German entries
(excluding German roots)



some statistics



most connected etymology entries

- English -ly (11156), un- (8822), non- (7872), -ness (7295), -er (6342), -ic (3357), -like, -able, -less, -y

- Hungarian -ok-,-ek-,-k-

- Italian -mente, -ità

- Finnish -sti

- German Shule, Haus, Stein, Holz, Sprache

- English water, back, head, work, wood

the idea

Create:

- database of etymological relationships exportable into Wikidata
  (wordA etymologicallyderivesFrom wordB)
- tool to visually explore etymological relationships between words
  - general user interested in etymologies
  - language learners

    (“meaningful learning […]: connected to prior learning […] more highly retainable and generalisable”*)
  - polyglots
  - discovering new words

* Herbert D. Pierson in “Using etymology in the classroom”

future steps

- improve parser
- improve visualisation: e.g., filter data
- integrate into Wikimedia:
    - export into Wikidata (after the WikibaseLexeme data model has been implemented)?
    - spot inconsistent/redundant/absent etymologies in Wiktionary
    - visualisations?
    - … ideas?

```
select count(?source) {
    ?source rdf:type ontolex:LexicalEntry .
}
```

```
select (count(distinct ?iso) as ?c){
    ?source ontolex:canonicalForm ?form.
    ?form rdfs:label ?label .
    bind (lang(?label) as ?iso)
}
```

```
select count(?source) {
    ?source rdf:type dbetym:EtymologyEntry .
    filter regex (str(?source),“_1_”)
}
```

```
select ?targetiso  (count (?source) as ?s) {
     ?source rdf:type dbetym:EtymologyEntry .
     ?source dbetym:etymologicallyRelatedTo ?target .
     ?source rdfs:label ?sourcelabel .
     ?target rdfs:label ?targetlabel .
     bind (lang(?sourcelabel) as ?sourceiso) .
     filter (?sourceiso = “deu”) .
     bind (lang(?targetlabel) as ?targetiso).
} order by desc(?s)
```

```
select ?target (count (distinct ?source) as ?s) {
         ?source rdf:type dbetym:EtymologyEntry .
         ?source dbetym:etymologicallyRelatedTo ?target .
} order by desc(?s) LIMIT 700
```