



Wiktionary Normalization of Translations and Morphological Information

Winston Wu and David Yarowsky

Center for Language and Speech Processing, Johns Hopkins University
{wswu,yarowsky}@jhu.edu <https://github.com/wswu/yawipa>



Yawipa is a **comprehensive** and **extensible** Wiktionary parsing framework. Help us develop Wiktionary parsers for your language!

What can Yawipa do?

Introduced in *Computational Etymology and Word Emergence* (Wu and Yarowsky, LREC 2020), Yawipa has comprehensive coverage of the English Wiktionary and partial support for several other editions. Compared to existing parsers, Yawipa aims to not only parse structured data (encoded as Wiktionary templates), but also information encoded as *unstructured, free-form text*. Yawipa takes the Wiktionary XML dump and outputs an easily-processable tabular format. Here is a sample of interesting data that Yawipa can extract:

The Standard Stuff

Part of speech, pronunciations, translations, cognates, derived terms, related terms, synonyms, antonyms, alternative forms, hyponyms, inflections, and much more!

Pronunciations

Extracts and normalizes IPA, phonemic pronunciation, dialectal variation, rhymes and hyphenation; useful for speech research.

IPA	/ˈnɒlɪdʒ/	variant=RP
IPA	/ˈnɑlɪdʒ/	variant=GA
enPR	nɒl'ij	variant=GA
enPR	nɒl'ij	variant=obsolete
IPA	/ˈnoʊlɪdʒ/	variant=obsolete
audio	en-us-knowledge.ogg	Audio (US)
rhymes	nɪdʒ	
hyphenation	know-ledge	caption=Hyphenation UK
hyphenation	knowl-edge	caption=US

Etymology

Etymology data is a mixture of Wiktionary templates interspersed in free-form text, comprising 489 relations in the English edition.

Text (Generated HTML)	Wiki Markup
From Middle English ethymologie, from Old French ethimologie, from Latin etymologia, from Ancient Greek ἔτυμολογία (etumología), from ἔτυμον (étumon, "true sense") and -λογία (-logía, "study of"), from λόγος (lógos, "word; explanation").	From {{inh en enm ethymologie}}, from {{der en fro ethimologie}}, from {{der en la etymologia}}, from {{der en grc ἔτυμολογία}}, from {{m grc ἔτυμον true sense}} and {{m grc -λογία study of}}, from {{m grc λόγος word; explanation}}.

Translations from Definitions

Translations of words may be listed in the Definition section of an entry but are not explicitly marked as translations. We use heuristic text processing to extract lexical translations.

Interlingua [edit]

Adjective [edit]

car (comparative plus car, superlative le plus car)

- dear; beloved; cherished
- expensive

Irish [edit]

ina	car	def tr	Adjective	dear
ina	car	def tr	Adjective	beloved
ina	car	def tr	Adjective	cherished
ina	car	def tr	Adjective	expensive
gle	car	def tr	Verb	love
gle	car	def tr	Verb	be devoted to

Verb [edit]

car (present analytic carann,

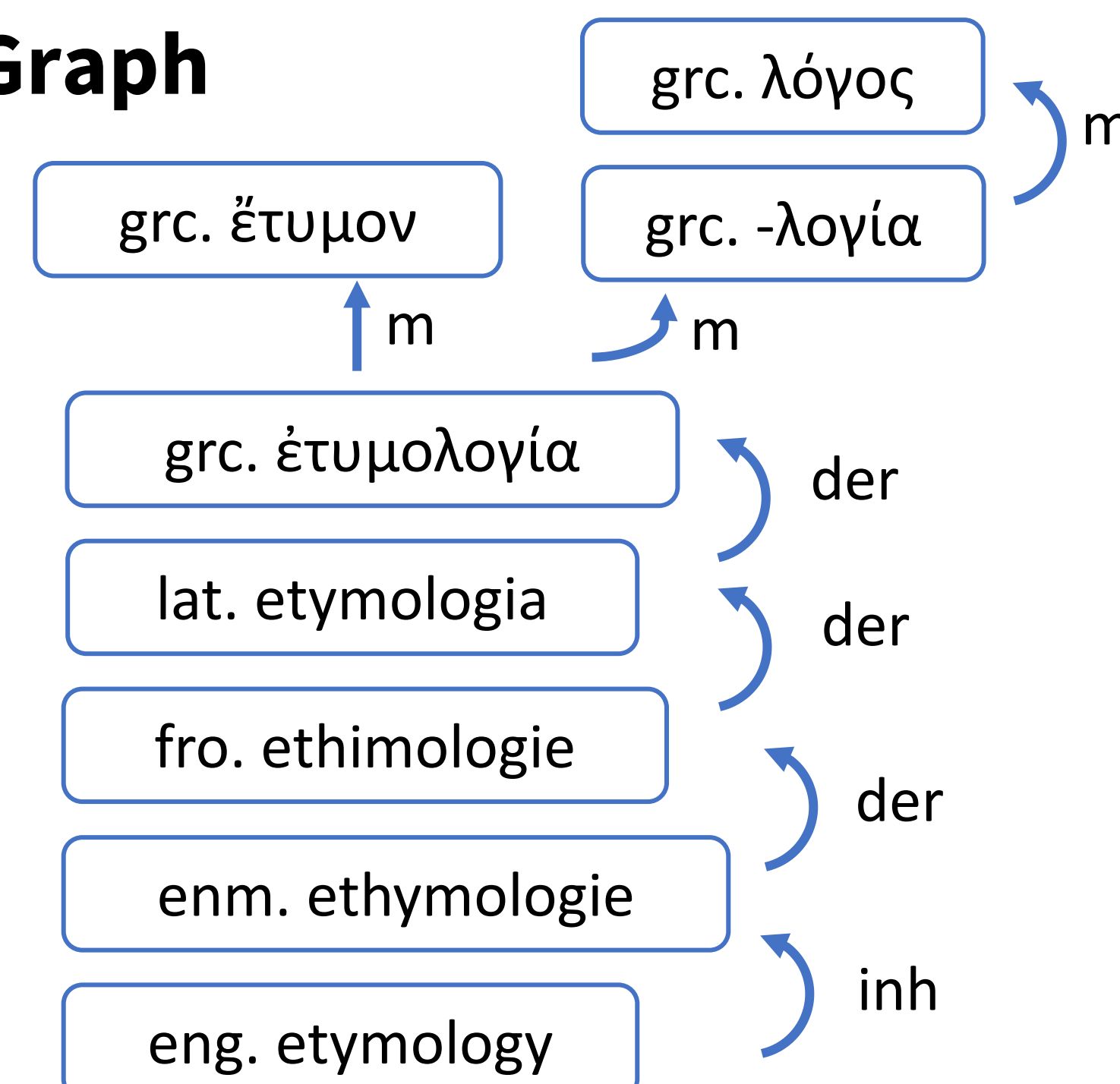
- to love
- be devoted to

Translations from Etymology Glosses

{{compound|de|Zeit|t1=time|Geist|t2=spirit}}

Another rich but overlooked source of translations. We augment Yawipa to extract translations from etymology glosses, adding 300K new translations.

Etymology Graph



Morphological "Form-Of" Relations

```
{{abbreviation of|en|caterpillar}}
{{alternative form of|enm|bouk}}
{{inflection of|fr|pondre||3|s|pres|indc}}
{{inflection of|la|piō||2|s|pres|actv|subj}}
{{nonstandard spelling of|cmn|sc=Latn|piē}}
```

We extend Yawipa's functionality to extract 4M instances of 168 relations, useful for computational morphology research.

Typo Detection

```
{{suffix|lv|afrikanis|iētis|gloss1=African}}
```

Gloss indices without a corresponding argument indicate a typo (the annotator typed '1' twice). Using this method, we identify a handful of such typos in the English Wiktionary.

Word Form Generation Experiments

Using the data we extracted, we train multilingual neural character seq2seq models to generate new word forms for the following word formation processes:

Contraction: I am -> I'm, not -> -n't

Clipping: mathematics -> math, telephone -> phone

Eye Dialect: after -> aftuh, joking -> jokin'

Experiment	Input Format	Output Format	Luong Attn		Copy Attn	
			1-best	5-best	1-best	5-best
Clipping	ht k a p a b	k a p	.25 (2.5)	.29 (2.0)	.38 (2.1)	.49 (1.5)
Contraction	en p a r e n t s ' r e n t s	r e n t s	.35 (1.7)	.49 (1.2)	.39 (1.5)	.54 (0.9)
Eye Dialect	t w e n t y	t w e n n y	.32 (1.6)	.42 (1.1)	.39 (1.5)	.48 (1.0)

Metrics: accuracy and (mean character edit distance)

Experimenting with two forms of attention, we find copy attention works better than a standard attention due to the nature of this generation task.