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.

**Etymology**

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

**Translations from Definitions**

Translations of words may be explicitly marked as translations. We use heuristic text processing to extract lexical translations.

**Translations from Etymology Glosses**

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

**Morphological “Form-Of” Relations**

{[abbreviation|en|caterpillar]}  
{[alternative form|en|bouk]}  
{[inflection of|fr|pondre]|3|pres|indc]}  
{[inflection of|la|pl|2|pres|actv|subj]}  
{[nonstandard spelling|cmn|sc|Latin|pl|3|]}  

We extend Yawipa’s functionality to extract 4M instances of unstructured, free-form text.

**Typo Detection**

{[suffix|lv|afrikans|lets|gloss=1|African]}  

Gloss indices without a corresponding argument indicate a typo (the annotator typed ‘i’ 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 -> aftu, joking -> jokin’

**Etymology Graph**

Using this method, we identify a handful of such typos in the English Wiktionary.

**Typo Detection**

{[suffix|lv|afrikans|lets|gloss=1|African]}  

Gloss indices without a corresponding argument indicate a typo (the annotator typed ‘i’ 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 -> aftu, joking -> jokin’

**Etymology Graph**

Using this method, we identify a handful of such typos in the English Wiktionary.

**Typo Detection**

{[suffix|lv|afrikans|lets|gloss=1|African]}  

Gloss indices without a corresponding argument indicate a typo (the annotator typed ‘i’ 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 -> aftu, joking -> jokin’

**Etymology Graph**

Using this method, we identify a handful of such typos in the English Wiktionary.

**Typo Detection**

{[suffix|lv|afrikans|lets|gloss=1|African]}  

Gloss indices without a corresponding argument indicate a typo (the annotator typed ‘i’ 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 -> aftu, joking -> jokin’

**Etymology Graph**

Using this method, we identify a handful of such typos in the English Wiktionary.

**Typo Detection**

{[suffix|lv|afrikans|lets|gloss=1|African]}  

Gloss indices without a corresponding argument indicate a typo (the annotator typed ‘i’ 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 -> aftu, joking -> jokin’

**Etymology Graph**

Using this method, we identify a handful of such typos in the English Wiktionary.

**Typo Detection**

{[suffix|lv|afrikans|lets|gloss=1|African]}  

Gloss indices without a corresponding argument indicate a typo (the annotator typed ‘i’ 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 -> aftu, joking -> jokin’

**Etymology Graph**

Using this method, we identify a handful of such typos in the English Wiktionary.

**Typo Detection**

{[suffix|lv|afrikans|lets|gloss=1|African]}  

Gloss indices without a corresponding argument indicate a typo (the annotator typed ‘i’ 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 -> aftu, joking -> jokin’

**Etymology Graph**

Using this method, we identify a handful of such typos in the English Wiktionary.

**Typo Detection**

{[suffix|lv|afrikans|lets|gloss=1|African]}  

Gloss indices without a corresponding argument indicate a typo (the annotator typed ‘i’ 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 -> aftu, joking -> jokin’

**Etymology Graph**

Using this method, we identify a handful of such typos in the English Wiktionary.

**Typo Detection**

{[suffix|lv|afrikans|lets|gloss=1|African]}  

Gloss indices without a corresponding argument indicate a typo (the annotator typed ‘i’ 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 -> aftu, joking -> jokin’

**Etymology Graph**

Using this method, we identify a handful of such typos in the English Wiktionary.

**Typo Detection**

{[suffix|lv|afrikans|lets|gloss=1|African]}  

Gloss indices without a corresponding argument indicate a typo (the annotator typed ‘i’ 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 -> aftu, joking -> jokin’

**Etymology Graph**

Using this method, we identify a handful of such typos in the English Wiktionary.

**Typo Detection**

{[suffix|lv|afrikans|lets|gloss=1|African]}  

Gloss indices without a corresponding argument indicate a typo (the annotator typed ‘i’ 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 -> aftu, joking -> jokin’

**Etymology Graph**

Using this method, we identify a handful of such typos in the English Wiktionary.

**Typo Detection**

{[suffix|lv|afrikans|lets|gloss=1|African]}  

Gloss indices without a corresponding argument indicate a typo (the annotator typed ‘i’ 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 -> aftu, joking -> jokin’

**Etymology Graph**

Using this method, we identify a handful of such typos in the English Wiktionary.