Enriching Multiword Terms in Wiktionary with Pronunciation Information

Lenka Bajčetić, Thierry Declerck, Gilles Sérasset
Innovation Center of the School of Electrical Engineering in Belgrade, DFKI GmbH, Université Grenoble Alpes

MWE 2023 : 19th Workshop on Multiword Expressions

Wiktionary as a source

- Free and open source
- Collaborative, constantly updated
- Rich information: categories
- Frequently used

English Wiktionary Statistics

711,641 Lemmas

75,401 Lemmas with IPA

157,883 Multiword Expressions

? Multiword Expressions with IPA

English Wiktionary Statistics

711,641 Lemmas

75,401 Lemmas with IPA

157,883 Multiword Expressions

6,767 Multiword Expressions with IPA

Data extraction

- Parsing the dump (fast, but incomplete)
- Wiktionary API (complete but very slow)
- DBnary (fast, and can be modified)

Data extraction

- Parsing the dump (fast, but incomplete)
- Wiktionary API (complete, but very slow)
- **DBnary** (fast, and can be **modified**)

Heteronyms

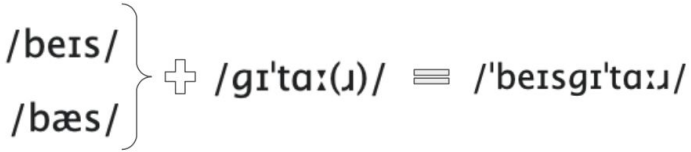
- Require disambiguation
- Can make use of Wiktionary “etymology” and “derived terms” sections

Current results

- MWE with IPA gold standard
- Added categories to DBnary
- Newly produced pronunciations added to Wiktionary

Current results

- MWE with IPA gold standard
- Added categories to DBnary
- Newly produced pronunciations added to Wiktionary
- Evaluation in progress: suprasegmentals



Thank you

Do you have any questions?

lenka.bajcetic@ic.etf.bg.ac.rs
declerck@dfki.de
Gilles.Serasset@univ-grenoble-alpes.fr

Work supported by the NexusLinguarum COST action (https://nexuslinguarum.eu/) and by the European LT-Bridge project (https://lt-bridge.eu/)