

New Corpora and Languages in Sketch Engine

Jan Michelfeit

Lexical  Computing

jan.michelfeit@sketchengine.co.uk

7th Sketch Engine Workshop
Portorož, May 23, 2016

New corpora

- Early English Books Online – texts from 1473 to 1820 (820M)
- Igbo WaC (330k)
- Parsed DeWaC – part of German Web Corpus, word sketches from dependencies (750M)
- Yoruba WaC (2.8M)
- Urdu WaC (53M)
- Nynorskorpuset – fiction, newspaper texts, journal articles, textbook texts, religious texts. . . (74M)
- Frantext – 500 works of French literature covering the period from the 18th to 20th century, copyright-free part only (15.5M)
- Mongolian Web Texts 2016 (6M)
- Latvian web 2014 – crawled by SpiderLing in 2014, tagged recently (530M)

- Links to audio clips in the spoken parts of British National Corpus

- Longest Commonest Match (LCM) finally compiled for (almost) all non-legacy corpora
- Human-readable corpus names:
enTenTen13 → *English Web 2013*
- Human-readable names of grammatical relations:
adj subject of → *adjective predicates of "test"*

New text processing pipeline for English

- character normalization
- new tagger model
- tokenizer: **facebook.com**, the **70s**, **mid-20th** century, ordinal numbers
- lemmatizer: **[url]**, **[number]**, 'd as **would**, guessing lemma for unknown plurals
- possessives retokenized: **Valentine's** as a single token, lemma **Valentine**, tags ending in Z (**NNZ**, **NNSZ**...)

New text processing pipeline for English

- character normalization
- new tagger model
- tokenizer: **facebook.com**, the **70s**, **mid-20th** century, ordinal numbers
- lemmatizer: **[url]**, **[number]**, 'd as **would**, guessing lemma for unknown plurals
- possessives retokenized: **Valentine's** as a single token, lemma **Valentine**, tags ending in Z (**NNZ**, **NNSZ**...)
- re-processed corpora: English Web 2013, BNC, Feed Corpus v6

New pipelines for other languages

- Latvian: already available for user corpora
- French: better clitic tokenization? (personal pronouns)
- German: better lemma guessing from suffixes
- Dutch
- Italian
- Serbian/Croatian/Bosnian
- ...

That's all

- Thanks for your attention. Questions?