GDEX for mere mortals

Jan/Honza Michelfeit

jan.michelfeit@sketchengine.co.uk

5th Sketch Engine Workshop
Bolzano, July 14, 2014

A short overview

- GDEX = Good Dictionary EXamples

A short overview

- GDEX = Good Dictionary EXamples
- EXamples = example sentences

A short overview

- GDEX = Good Dictionary EXamples
- EXamples = example sentences
- Good = not bad

A bad sentence is

- too short or too long
  (a good indicator of other issues)
- too specific – names, numbers
  $\rightarrow$ numerals, uppercase letters, tagged proper names
- too vague – anaphors
  $\rightarrow$ pronoun greylist, tagged pronouns

A bad sentence is

- too short or too long
  (a good indicator of other issues)
- too specific – names, numbers
  $\rightarrow$ numerals, uppercase letters, tagged proper names
- too vague – anaphors
  $\rightarrow$ pronoun greylist, tagged pronouns

. . . comprehensibility out of context

A bad sentence has

Noise:

- common web corpus noise
  → correct sentence structure and optimal length
  → number of symbols and punctuation marks
- typos, slang and other non-words
  → minimum word frequency limit
- profanities, sensitive topics. . .
  → blacklist or greylist

Current achievements

- GDEX for Slovene (Kosem, Husak, McCarthy, 2011)
- automatic collocation dictionaries
- simple english for language learners

The old format

- (let's take a look at the configuration)
- complicated syntax
- web interface not very flexible
- very addition-oriented (weighted average)

The old format

- (let's take a look at the configuration)
- complicated syntax
- web interface not very flexible
- very addition-oriented (weighted average)
- $\rightarrow$ simplify, make human-readable/writable

New configuration

```
formula: >
 (50 * is_whole_sentence()
  * blacklist(words, illegal_chars)
  * blacklist(lemmas, parsnips)
  * (min([word_frequency(w) for w in words]) > 3)
 + 20 * optimal_interval(length, 10, 14)
 + 15 * greylist(words, rare_chars, 0.1)
 + 15 * greylist(tags, pronouns, 0.1)
 ) / 100
variables:
  illegal_chars: ([<|\]\[>/\\^@])
  rare_chars: ([A-Z0-9'.,!?)(;:-])
  pronouns: PRON.*
  parsnips: ^(arse, bollocks, tory, whig, booze)$
```

The new format

- a simple arithmetic expression in Python
- predefined variables like length, words, tags. . .
- regular expressions
- same expressive power as the old format

Current status

- deployed on beta.sketchengine.co.uk
- explore: less addition, more multiplication
- values can be pre-calculated