Expert features of Sketch Engine

Vojtěch Kovář

Lexical Computing

vojtech.kovar@sketchengine.co.uk

7th Sketch Engine Workshop
Portorož, May 23, 2016

Translate button

- Going from monolingual word sketch to bilingual
  - without having to specify the translation
- Needs proper set-up
  - technically a bit complicated
  - by now for Europarl 7 corpora
  - but more to come soon :)
  - as well as more compatible sketch grammars
- Available for parallel, but also for comparable corpora

Word sketch not structured by relations

- Switch show/hide gramrels in the left menu
- Global sorting according to score/frequency
  - best collocations across relations on the top
  - different score computation, does not take relations into account

Smart working with n-grams

- On the word list page
- Filtering n-grams
  - filter n-grams output for specific word(s)
- Nesting n-grams
  - "at the end", "end of the", "of the day", "at the end of the day"
  - display all of them in one nest

Meta-data overviews

- Also on the word list page
- Search attribute contains also meta-data attributes
  - compatible with filtering options

Advanced features of CQL

- Queries based on thesaurus
  - search a word plus X most similar words
  - e.g. search for animals
  - `[lempos_lc~30"goat-n" & word!="[A-Z].*"]`
- Word sketch based queries
  - sentences containing a particular collocation(s)
  - can use regular expressions
  - `[ws("woman-n","modifiers of \"%w\"","good-.*-j")]`
- Similar term-based queries

Creating subcorpora without meta-data

- Standard way: from meta-data
- Recently added: from concordance
  - documents/paragraphs/sentences containing ...
  - ... a particular word
  - ... a particular phrase
  - ... a particular meta-tag
  - ... a particular phenomenon specified by CQL
  - ... a particular word, or one of 10 most similar
  - ... a particular word sketch collocation
  - ...
- Simple and transparent solution to the lack of good meta-data in web corpora?

New Access to Sketch Engine API

- Now on a dedicated server
  - `https://api.sketchengine.co.uk`
- Simple authentication separate from the main installation
  - generate an API key on the main installation
  - connect to the server without any authentication
  - put username and API key as parameters into the URL

```
https:
//api.sketchengine.co.uk/bonito/run.cgi/wsketch?
corpname=susanne;lemma=man;format=json;username=
xkovar3;api_key=PWPZI3I5IKW0MCLLV0X2V5PIGGDN5RI6
```

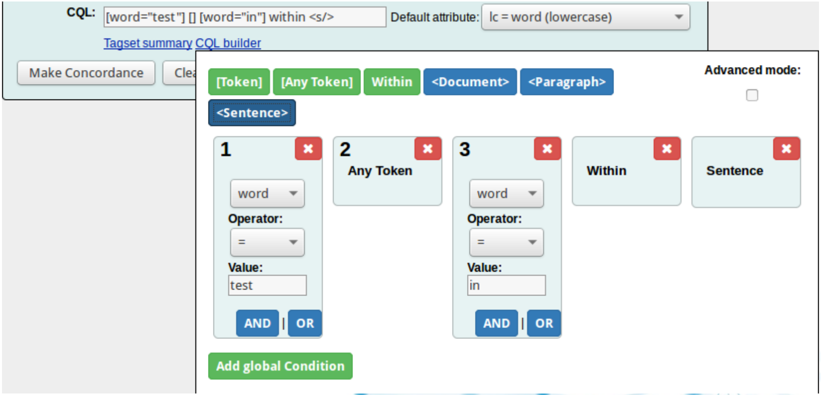
Small new improvements in the interface

- Concordance → Search
- Longest-commonest match in word sketches
  - shows typical usage of the collocation, if available
  - longest context that covers 1/6 or more of the concordance lines
  - and has at least 5 hits
- French and Spanish localization (Turkish localization in process)
- Word sketch score of grammatical relation
  - 100 * no_hits / headword_freq

Soon on your monitors

- Re-working the word list functionality
  - there is a lot of options to combine...
  - ... and people get typically lost
  - divide into separate functions
- User interface facelift
- Minor re-organizations of the interface
  - mainly to make it more intuitive
- Graphical interface for creating CQL queries

Soon on your monitors



CQL: [word="test"] [] [word="in"] within <s/>

Default attribute: lc = word (lowercase)

Advanced mode: □

1: word, Operator: =, Value: test

2: Any Token

3: word, Operator: =, Value: in

Within

Sentence