Virtual corpora and subcorpora

Pavel Rychlý

pavel.rychly@sketchengine.co.uk

4th Sketch Engine Workshop
Tallinn, October 16, 2013

Use case

Search in several corpora of the same language.

Use case

Search in several corpora of the same language.

- repeat the same query several times (in the respective corpus)
- create a new corpus via concatenating selected corpus source texts

Drawbacks of concatenated corpora

- multiple copies of selected corpora
- processing time

	size [M tokens]	storage	compilation time
Brown	1.2	76 MB	1m
BNC	112.2	2 GB	45m
enTenTen12	12,970.0	293 GB	65h 30m

Virtual corpora

- creating a bigger corpus from a set of corpora
- works on already compiled corpora (no text sources)
- handles also subsets of corpora

Corpus configuration file – estenten11

New `VIRTUAL` option with a path to the specification file.

```
NAME     "esTenTen11 (European + American, Freeling)"
PATH     "/corpora/manatee/estenten11_freeling/"
VIRTUAL  "/corpora/virtdef/estenten11_freeling"
ENCODING "utf-8"

LANGUAGE "Spanish"
INFOHREF "http://www.sketchengine.co.uk/documentation/wiki/

ATTRIBUTE word {
    TYPE "FD_FGD"
}
```

Specification file

- list of corpora
- list of token intervals

```
=eseutenten11_freeling
0,2341159406

=esamtenten11_freeling
0,$
```

Results

Virtual corpus:

- estenten11 = esEUtenten11 + esAMtenten11
- 24 GB instead of 280 GB
- 17 hours instead of 60 hours

Use case II

Search only part of a corpus.

Use case II

Search only part of a corpus.

- use the *Text Type* section of the query form
- use one of predefined subcorpora
- create own subcorpus

Subcorpora

- limit *query search* to some parts of the whole corpus
- compute hits per million on these parts
- but the context is from original corpus
  - collocations, frequecies, visualization are based on whole corpus

Subcorpora

- limit *query search* to some parts of the whole corpus
- compute hits per million on these parts
- but the context is from original corpus
  - collocations, frequecies, visualization are based on whole corpus
- create a virtual corpus