CQP tutorial

Judith Degen

March 16, 2009

For a more complete tutorial, see Stefan Evert's CQP tutorial ((1))

1 Getting started

To start:

```
$ cqp -e
```

To see available corpora:

```
> show corpora
```

Show information about a given corpus:

```
> info BNC-XML
```

Activate corpus:

```
> BNC-XML
```

Show corpus attributes:

```
> show cd
```

Search for words (via regular expressions) and sort:

```
> "spe(ech|aks?(ing)?)"
> sort by word
```

Set context to 8 words preceding the target, 2 sentences following the target, 1 sentence pre- and post-target:

```
> set lc 8 words
> set rc 2s
> set c s
```

Redisplay matches:

```
> cat
```

Display or hide POS and lemma annotation:

```
> show +pos
> show +lemma
> show -pos -lemma
```

Search by lemma:

```
> [lemma = "speak_VERB"]
> [lemma = "speech_SUBST"]
> [lemma = "(speak_VERB|speech_SUBST)"]
```

See size of last query:

```
> size Last
```

Show structural attributes (shown as XML tags):

```
> show +s
```

Create .cqprc file with favorite settings:

```
set ProgressBar on;
set HistoryFile "/tmp/cqphistory.jdegen";
set WriteHistory yes;
set c s;
```

Searching for POS information:

```
> "work"
> [word="work" & pos="N.*"]
> [word="work" & pos="V.*"]
> [word="work" & pos !="V.*"]
```

Use `/codist[]` macro to get frequency distributions of POS-tags/lemmas over a given word:

```
> /codist["work",  pos]
> /codist[lemma, "speak_VERB", word]
```

Search for sequences, search within a context:

```
> [lemma="work_VERB"][]*[word="day"]
> [lemma="work_VERB"][]*[word="day"] within s
> [lemma="work_VERB"][]*[word="day"] within 2 words
> [lemma="work_VERB"][]{2}[word="day"]
```

Count:

```
> count by word
> count by lemma
```

Set frequency thresholds:

```
> [pos="VVB" & word = "w.*"]
> count by lemma cut 50
```

Save query results:

```
> Some = [word = "some" %c] [pos="NN2*"]
> set DataDirectory "."
> BNC-XML
> save Some
> cat Some > "some.txt"
> cat Some > "| gzip > some.txt.gz"
> sort Some by word
```

Anchor points:

```
> A = [pos="(AT.*|DT.*)"] @[pos="AJ.*" & word="f.*"] [pos="N.*"]
> sort by word
```

Display corpus positions of anchor points in tabular format:

```
> dump A
> dump A 10 20
```

Frequency distributions:

```
> group A matchend word by target word cut 100
> group A match word by target lemma cut 100
```

Reduce data randomly:

```
> reduce A to 10%
```

References

[1] S. Evert: *The CQP Query Language Tutorial* (2005).