# New features in CQL

Miloš Jakubíček



milos.jakubicek@sketchengine.co.uk

6<sup>th</sup> Sketch Engine Workshop Herstmonceux, August 10, 2015

# CQL

- $\bullet \ \mathsf{CQL} = \mathsf{Corpus} \ \mathsf{Query} \ \mathsf{Language} \\$
- all corpus queries are translated into it
- originated 1994 in Stuttgart (Christ and Schulze)
- adopted by Manatee and extended since then

#### Recent development

- general NOT operator, NOT within, NOT containing
- word sketch queries, thesaurus queries, swap and ccoll
- different attributes in global conditions
- regular expression optimization

#### NOT (!) operator

- attribute level: [word!="castle"]
- within/containing: within !, containing !
  - NOT means corpus complement part here
  - often not the expected semantics
- new: !within, !containing
  - <doc year="2010"> ! containing [word="castle"]
- new: general ! <query>
  - corpus complement meaning

#### Global conditions

- at the end of the query:
  - 1: [lemma="alt"] 2: [lemma="Schloss"] & 1.case = 2.case (German *old castle* agreeing in case)
- now available for different attributes: 1: [lemma="alt"]
   2: [tag="NOUN"] & 1.head = 2.id
   (dependency relation)

### Queries on word sketches

- three new operators: ws, swap, ccoll
  - all collocations by default labeled with 1 (as if produced by 1: [word="<collocation>"])
- modifiers of test (noun):
   [ws("test-n", "modifier", ".\*")]
- ...with test being in dative:
   [ws("test-n", "modifier", ".\*") & tag="\*.DAT.\*"]
- ...with collocates being in dative:
  [swap(1, ws("test-n", "modifier", ".\*")) &
  tag="\*.DAT.\*"]
- modifiers of test and experiment (nouns) with the same tag:
   [ws("test-n", "modifier", ".\*") | ccoll(1, 2,
   ws("experiment-n", "modifier", ".\*"))] & 1.tag = 2.tag

### Queries on thesaurus

- new tilda operator (~): [WSATTR~NUMBER"word"]
- [lempos~10"test-n"]

  test and its 10 top thesaurus entries
- [lempos~"test-n"]
   NUMBER is log<sub>10</sub>(freq(word)) by default
- ~"test-n" automatically choose word sketch attribute

# Regular expression optimization

- queries with regular expression take more time to execute
  - one must find matching lexicon entries first
- three optimization schemes
  - simple OR optimization
  - prefix optimization
  - n-gram prefetching

- simple OR optimization
  [word="a|b|c"] ⇒ [word="a" | word="b" | word="c"]
- prefix optimization
   [word="re.\*"] ⇒ lookup sorted words between re and rd
- n-gram prefetching
  - preindex all character one-, bi-, trigrams
  - more in Optimization of Regular Expression Evaluation within the Manatee Corpus Management System (Jakubíček and Rychlý, 2014)

# **Evaluation**

corpus	query	#RE w/o	#RE w/	time w/o	time w/	S
czTenTen12	[word=".*ější"]	18,978,703	41,426	10.030	0.341	29.4
	[lemma=".*strč.*"]	14,151,454	888	6.601	0.066	100.0
	[tag="k1.*c4.*"]	1,357	251	0.058	0.049	1.2
	[word="[sz]p.*"]	18,978,703	115,347	10.023	0.698	14.4
enTenTen12	[word=".*ing"]	27,894,538	913,004	21.931	2.768	7.9
	[lemma=".*ten.*"]	26,426,200	195,758	22.163	1.218	18.2
	[word="pre.*ed"]	80,054	7,553	0.294	0.178	1.7
	[word="pr[oe].*"]	251,924	198,297	1.329	0.920	1.4
	[word=".*[dt]"]	27,894,538	3,466,379	41.100	8.538	4.8
	[tag="N.*"]	60	5	0.056	0.048	1.2
jpTenTen11	[word=".*5.*"]	13,844,200	30,160	8.182	0.364	22.4
	[lemma=".*ア.*ス"]	13,303,479	69,228	8.388	0.450	18.6
	[word="ンテ.*"]	17,078	17,077	0.199	0.178	1.12

#### Technical notes

- to build the n-gram index, use:
  - mkregexattr <CORPUS> <ATTRIBUTE>
- part of Manatee version 2.111, automatically called by encodevert for attributes with lexicon exceeding 10,000 items

### Conclusions

CQL still under development and being extended continuously.