QUICK REFERENCE GUIDE TO CQPweb SEARCHES AT GU

Sakol Suethanapornkul and Amir Zeldes
Department of Linguistics, Georgetown University
January 21, 2015

This reference outlines a brief overview of the most commonly searched features of CQP Query Syntax in the CQPweb interface at Georgetown University. Query expressions are entered in a query field on the “Standard query” page available from the menu on the left in each corpus. For a list of corpora, see http://corpling.uis.georgetown.edu/cqp/.

Basic word searches – Simple query mode

By default, the search query mode is set to **CQP syntax**. For basic word form searches, set the Query mode to **Simple query** (ignore case)

- This reference primarily deals with CQP mode. In Simple query, to search for a word or sequence of words, enter the word into the query field and click Start Query button, for example: dry or even though. For more information on simple query syntax, click the link next to the query mode box.

Running CQP mode queries

Lemma and part of speech (POS) searches

- Use `[word="word"]` to search for a particular word.
  `[word="take"]` → find all instances of exactly the word "take".
- Use `[lemma="word"]` to search for a lemma, including inflected forms.
  `[lemma="dry"]` → `dry`, `dried`, `drying`, `dries`…
- To search for part of speech tags, use `[pos="POS"]`
  `[pos="JJ"]` → find all adjectives. Most (but not all) corpora use the extended Penn Treebank tag set – for a list see here: http://corpling.uis.georgetown.edu/ptb_tags.html or the list at the end of this manual.

Searches are case-sensitive. For example, `[word="Take"]` and `[word="take"]` produce different results.

Operators in searches

- Use operators to search for patterns:

`.` The dot operator matches any character:
`d.g` → `dog, dig, dug`

`*` matches the preceding character zero or more times.
`of*` → `o, of, off, offff...`

`+` matches the preceding character once or more:
`of+` → `of, off, offf, offff...`

`?` makes the preceding character optional:
`[word = "honou?r"]` → `honor, honour`

`(|)` searches for two alternative forms:

`[word="be(tter|st)"]` → `better, best`

`[lemma="(slew|slayed)"]` → find `slayed` or `slew`

`[]` defines a range of characters.

`[aeiou]` → any vowel, e.g. [word="d[aeiou]g"]

`[A-Z]` → similarly, a capital letter A to Z

`[0-9]+` → a sequence of numbers (one or more, using the + from above)

For example, you can combine these options like this:

`[word="[A-Za-z]+-[A-Za-z]+"]` → find hyphenated compounds

`[^]` defines the opposite of a range.

`[^aeiou]` → anything but a vowel

`[^a-z]+` → a string of only non-lower case characters

You can use these like this:

`[word="[^aeiou]+"]` → find a word that does not begin with a vowel

`{n,m}` specifies a number range for repetitions

`[word="a{3,4}"]` → find `aaa` or `aaaa`.

- To treat operators as a real character, use `\` in front of the operator.
  `[word="\?"]` → find a `"?"` in the text.

Combining and negating annotations

- Combine search terms using &:
  `[pos ="JJ" & lemma = "dry"]` → find all instances of "dry" as an adjective.
- Use `!=` for a negative match:
  `[pos="JJ" & word !=".*able"]` → find adjectives that **don't** end with -*able*.
  `[pos= "NNS" & word!=".*s"]` → find irregular noun plurals.

Word sequence searches

- Combine search terms to look for a string of words.
  `[word="a"][word="lot"][word="of"]` → find the phrase *`a lot of`*.
- Use operators on tokens or annotations to search for patterns. The same operators that apply to characters can also be placed after each word:
  `[word="a"][pos="JJ"]*[word="lot"][word="of"]` → find *`a (ADJ) lot of`* with any number of adjective (*a great whole lot of…*)
  `[pos="JJ.*"]{2,4}[pos="NN.*"]` → find 2 to 4 consecutive adjectives before a noun

Markup searches

- Use XML markup for searches in corpora that support markup. Some corpora have sentence segmentations:
  `<s> [word="[Nn]o"]` → find a sentence that begins with 'no'.
- Others have paragraphs (p) or other mark up:
  `<p>[pos="V.*"]` → find a paragraph that begins with a verb.

Extended Penn Treebank part-of-speech (POS) tags

Tag	*Description*	*Example*
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	there [is]
FW	foreign word	d'hoevre
IN	preposition, subordinating conjunction	in, of, like
IN/that	that as subordinator	that
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NP	proper noun, singular	John
NPS	proper noun, plural	Americans
PDT	predeterminer	both [the boys]
POS	possessive ending	[friend]'s
PP	personal pronoun	I, he, it
PP$	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	[give] up
SENT	Sentence final punctuation	. ! ?
SYM	Symbol	/ = *
TO	infinitive or prepositon 'to'	to
UH	interjection	hey, huh, uh
VB	verb be, base form	be
VBD	verb be, past tense	was, were
VBG	verb be, gerund/present participle	being
VBN	verb be, past participle	been
VBP	verb be, sing. present, non-3rd	am, are
VBZ	verb be, 3rd person sing. present	is
VH	verb have, base form	have
VHD	verb have, past tense	had
VHG	verb have, gerund/present participle	having
VHN	verb have, past participle	had
VHP	verb have, sing. present, non-3d	have
VHZ	verb have, 3rd person sing. present	has

VV	verb, base form	take
VVD	verb, past tense	took
VVG	verb, gerund/present participle	taking
VVN	verb, past participle	taken
VVP	verb, sing. present, non-3d	take
VVZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when
``	Opening quotation marks	' "
"	Closing quotation marks	' "
(	Opening brackets	({
)	Closing brackets	) }
,	Comma	,
:	Other punctuation	- ; : -- ...