Manual for the CQPweb interface at CLUL

Manual 1.2

Version May, 2014

Michel Généreux, Iris Hendrickx, Amália Mendes

Centro de Linguística da Universidade de Lisboa
Complexo Interdisciplinar
Av. Prof. Gama Pinto, 2
1649-003 Lisboa - Portugal

Table of Contents

Preamble

This manual explains how to use the CQPweb interface to query corpora. The query language (*Simple Query Syntax)* is almost the same as for the BNCweb which is described in detail in Chapter 6 of Hoffmann, Sebastian et al. (2008), *Corpus Linguistics with BNCweb - a Practical Guide*. Frankfurt/Main: Peter Lang.

Several corpora are available on the CQPweb interface at CLUL. We present here the main features of the interface, common to all corpora.

A corpus manual is also available for most of the corpora on CQPweb at CLUL. Refer to these manuals to find out which levels of annotation can be queried and how.

1. Search for concordances of word forms

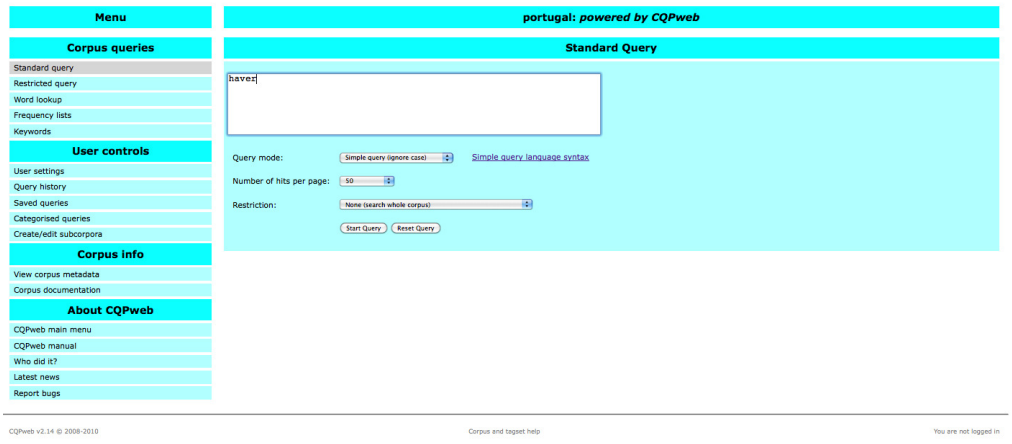
To conduct a simple query, go to the main page, without changing any option, insert a word or a sequence of words in the box and click on *Start Query*. At the top of the window with the results, there is information on the number of contexts, the number of texts in which the query occurs and information about the total corpus.

To make a new search, click GO (top right button).

To view information on the full text that matched the query for a particular concordance, click on the name in the left column "Filename". Any user can download its concordances by selecting “Download” on the drop-down list located at the top-right corner and click Go!.

To view a larger context of a particular concordance, click on the bold words on the intended line. You will see the words in a context of a few lines. In the top menu there is the possibility to enlarge the context, click on "More context". You can see the part-of-speech tags by clicking "Show tags".

WARNING: very common grammatical words such as "que " "de", "o ", can be queried, but the lookup takes time due to very high frequency of these forms in large corpora such as CRPC.



1.1 Regular expressions

A regular expression is a way of characterizing a string, you can view it as a pattern or a template in which you use wildcards to leave certain characters unspecified.

Wildcards	*example*	*matches with*
? a single arbitrary character	gat?	gato, gata
* zero or more characters	*mente	mente, absolutamente, provavelmente, etc.
+ one or more characters	+mente	absolutamente, provavelmente, etc. (but not: *mente*)

Simple Query Syntax uses a set of characters as meta-characters:

? * + [ ] ( ) { } , : @ / _ - < >

To query for the literal meaning of these characters, use a backslash in front.
E.g. to look for a question mark, type: *\?*

Query type	*example*	*matches with*
Alternatives: between square brackets	lind[o,a]	lindo, linda
Two alternatives followed by exactly 1 character	lind[o,a]?	lindos, lindas
Two alternatives followed by: 's' or nothing	lind[o,a][s,]	lindo, linda, lindos, lindas
Two alternatives followed by zero or more characters	lind[o,a]*	lindo, lindos, lindamente, lindinho, lindoso, lindano, etc.

1.2 Word sequences

You can also search for multiple words. Notice that:

- punctuation marks are split from words and are separate tokens
- special characters need a backslash
- you can combine + and * to define a sequence of arbitrary words in your query. E.g. the pattern +** represents a sequence of one to three tokens.

Query description	*example*	*matches with*
The word 'célebre' followed by the word 'jantar'	célebre jantar	célebre jantar
The word 'se' followed by an optional word and a comma	se * \,	se trata, se, se vê, se calhar, etc.
The preposition ‘de’ followed by the the word 'jantar', separated by a minimum of one and a maximum of three words	de +** jantar	de estar presente num jantar, de fazer um jantar, de nosso jantar, etc

1.3 Sort concordances

After searching for a word or expression, you can sort the concordances obtained: open the window New Query and click on Go!

By default, the concordances are sorted alphabetically by the first word on the right. You can change this option in “Position” and then click on "Update sort".

2. Main Left menu

2.1 Corpus Queries

- **Standard query** See section 1 above about standard searches.

- **Restricted queries** This enables you to search in a particular sub set of the corpus using metadata fields to restrict your query.
- **Word lookup** Use this option to get frequency information about a particular word. You can also use regular expressions or only specify the beginning or end of a word. When you click on a word in the result page, you will get a concordance list.

- **Frequency lists** Gives a list of all word forms or lemmas from the corpus and their frequency.

- **Key Words** This rather advanced option allows you to compare a query in a restricted sub corpus against the full corpus.

2.2 User controls

User controls are only available for registered users (the green version). This means essentially that unregistered users (the blue version) cannot *save* data (settings, queries and sub-corpora) on our server. However, they can *download* their results and benefit from exactly the same searching power available to registered users.

User settings Various user-oriented options.

Query History Shows all previously entered queries.

Saved queries When making a standard or restricted query, results can be saved. These saved queries are listed here. Registered users should keep the number of saved queries to a useful minimum by using the delete function.

Categorized queries The set of concordances obtained through a regular or restricted query can be organized using a set of labels applied to each individual context.

- Select the option "categorize" on the top right menu and click Go!

- Enter a name for the set of categories. For example, if you want to label each sense of a highly polysemous verb like "abater" (*move downwards / eliminate / negatively affect*) the set of values could be named "abater" or "verbpolysemy".

- Enter the names for each category. For example, considering the different senses of “abater”, the set could be: movement, movement_pronominal, psych, psych_pronominal, affect, affect_eliminate, subtract, etc.

- select the default value (for example, if the verb has a more frequent sense)

- click on Submit

The set of concordances will appear with a new column named 'Category' on the right, with the set of values to select. Two categories are automatically added to the set you have created: 'other' and 'unclear'.
After selecting a value for each context, select "save values and leave categorisation mode".

The set of categorised concordances remains available on the left menu. There are two interesting options under User Controls:
- add categories
- separate categories: this creates a separate list of concordances for each category, with information on the number of hits of each.

Create/edit sub corpora You can create separate sub corpora based on several criteria such as using the meta data from the corpus or using the matches from a query. For example, you can create a sub corpus containing only Portuguese news texts from the CRPC:
- select ‘corpus meta data’
- click Go!
- enter a name for this new subcorpus
- choose the text-type restrictions 'Portugal' and 'jornal'
- click on ‘Create subcorpus from selected categories’.

Next you can compile a frequency list for this sub corpus by clicking ‘Compile’ under Frequency Lists on the left Menu. This frequency list can be further inspected using the option " Frequency lists" in the main menu. Registered users should keep the number of saved subcorpora to a useful minimum by using the delete function.

3. Collocations

You can obtain additional collocation information for a retrieved word or lemma pattern from a standard or restricted query. Instead of choosing 'new query', choose 'collocations' from the menu drop box and click 'go'. Click on "Create collocation database” to get the list of words that co-occur with the retrieved word pattern.

On top, "Collocation controls", you can:
- change the statistical measure used (statistics: Mutual information, t-score, etc.) and compare the results
- change the distance between words (collocation window)

In "submit changed parameters", press Go!

By clicking on the frequency, you get the concordances in which the word you searched co-occurred.

Below is a screenshot for collocations for the word 'janela' in a search window of 3 words to the left and right using Log-likelihood as distance measure, and a frequency threshold of 5.

Collocation controls

Collocation based on:	Word form	Statistic:	Log-likelihood
Collocation window *from*:	3 to the Left	Collocation window *to*:	3 to the Right
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5
Filter results by:	specific collocate:	and/or tag: (none)	Submit changed parameters Go!

There are 7,462 different words in your collocation database for “[word="janela"%c]”. (Your query “janela” returned 5,283 matches in 2421 different texts) [34.006 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood value
1	pela	459,379	47.001	606	460	1991.248
2	uma	1,971,066	201.666	1,077	799	1882.461
3	ABERTA	22,282	2.28	238	206	1744.769
4	da	4,537,364	464.233	1,480	764	1433.505
5	à	10,957,196	1121.067	2,476	1075	1274.909
6	quarto	18,750	1.918	126	107	807.36
7	Abrir	15,686	1.605	99	89	622.026
8	Indiscreta	96	0.01	39	26	586.634
9	abriu	8,028	0.821	70	53	484.527
10	parapeito	158	0.016	35	31	475.353
11	vidro	3,971	0.406	58	52	460.675
12	para	2,629,825	269.066	666	504	418.273

4. Sentences and Noun Phrases

Version 2.2 of the CRPC has been tagged with Noun Phrases (NPs). You can query those NPs provided you use the CQP syntax. Here are a few examples:

All NPs: (this will take a very long time!)
/region[np];
<np> []* </np>;

NPs with exactly 3 words:
<np> []{3} </np>;

V at the start of a sentence:
<s> [pos = "V"];

V at the start of a sentence:
[(pos = "V") & lbound(s)];

V at the end of a sentence:
[pos = "V"] [pos = "PNT"]? </s>;

NP with at least 3 adjectives:
<np> []* ([pos="ADJ.*"] []*){3,} </np>;

Sentences that start and end with a NP:
<s><np>[]*</np> []* <np>[]*</np></s>;

CN that is not contained in a noun phrase:
[(pos = "CN") & !np];

Sequence of two singular nouns within the same NP:
[pos="CN"] []* [pos="CN"] within np;