ARCHER 3.2 in CQPweb: Hints and tips

This guide will be updated from time to time. Please send any suggestions for improvement to archer@manchester.ac.uk with the Subject line 'Hints and tips'. Last updated 28/04/2017 10:34.

The CQPweb interface

See Hardie (2013) for a full description of the software.

Searching

The default search is case-insensitive. Just type a word or phrase in the box and click **Start Query**. You can use wild cards either to represent characters in a word or to act as a whole word:

- **?** [= exactly one]
- ***** [= string of 0 or more]
- **+** [= string of 1 or more]

You can represent alternatives in various other ways, including separated by a vertical bar within round brackets:

(he|hee) (said|sayd|saide)
(should|shou'd|shd)

To find spelling variants before you construct a search, click **Frequency lists** at the left, sort alphabetically and start near your likeliest first variant. (The complete word list[1] in frequency order is also available on the ARCHER website.) Intuition and the *OED* may also suggest variant spellings. A later internet version of ARCHER will have spelling lemmatisation under modern British spelling.

You can restrict a query to particular varieties, genres and/or dates by clicking **Restricted query** at the left and checking the appropriate boxes; for quick selection of one period you can use the **Restriction** box in the centre instead. For repeated use you can **Create/edit subcorpora** of your own choice.

Inspecting the hits

When you see a list of hits, you can mouse-over or click on a filename to see the metadata, or click on the search term to see the wider context of that hit.

Context is limited for copyright reasons. In this version, there is context of up to 10 tokens left and right in the list of hits, up to 50 tokens L & R in the extended context (from which you can copy and paste), but for technical reasons only up to 10 tokens L & R in a download file. We hope to increase the latter in due course. For unlimited context you must visit a consortium university.

To see how variants are distributed (if your search allowed for them), or simply to home in on one variant, choose **Frequency breakdown** from the dropdown list at top right and click **Go!**. That dropdown list also offers other options for processing a list of hits, including **Thin**, **Sort**, **Distribution** and **Download** (among others).

[1] http://www.humanities.manchester.ac.uk/medialibrary/llc/files/ARCHER/wordlist_3-2.txt

ARCHER_untagged

This is an interim version to make ARCHER available to users without further delay.

Filenames

Filenames are as described on the ARCHER website,[2] following the formula **nnnnabcd_gpv**, where **nnnn** = year, **abcd** = abbreviation of author's surname, **g** = genre, **p** = period, **v** = variety. The one exception is where a hyphen should appear in a filename, which in CQPweb becomes an underscore, e.g. 1697pix__d2b, 1675br___s2b rather than 1697pix-_d2b, 1675br--_s2b. Please continue to use the form with hyphen(s) when referencing.

Mark-up

The full non-linguistic mark-up of the corpus is contained in TEI headers and XML tags, available only in the XML files at a consortium university. The present CQPweb version has the essential metadata at the level of the file, including date, author, title of work, publication data, sex of author, but not at a lower level, e.g. sex of speaker in drama. There is no linguistic mark-up in this version.

Tokenisation

Tokenisation of a corpus allows for separate indexing and searching of punctuation marks and other items, regardless of whether they are separated by whitespace in conventional orthography.

In this version there is no tokenisation of the possessive *'s* morpheme, the contracted negative *n't*, or contracted verbs like *'ve,'d,'s*.

Punctuation **is** generally tokenised, separated by whitespace and counted in the overall word count. After major punctuation the next word starts a new line in the full context display.

There is only partial discrimination of apostrophes from single quotation marks. The <'> character in mid-word is treated as an apostrophe: <boy's>,<don't>. If adjacent to white space or punctuation, however, it is tokenised separately as if it were a quotation mark, with a space on either side. In many cases this is appropriate, but not always: <ma '> ['Mama'] and <' tis, ' twas> should really have apostrophes, while only 2 out of 7 examples of <boys '> actually involve a closing quote.

Likewise there is only partial discrimination of full stops that belong to abbreviations from those used as sentence punctuation. However, some common abbreviations have been retained as a single token: <Mr. Mrs. St. Dr. Co. Capt. Jan. Dec. Yes. Feb. Fig. Gen. Col. Rev. Nov. Ltd. Sr. Oct. Sept. Esq. Aug. Reg. Mar. Hon. Wm. Ch. Ld. Rep. Ed. Genl.>

More elaborate searches

Proximity searches

Proximity queries are discussed in Hoffmann et al. (2008: 114-16). Examples of proximity searches:

expect* >>6>> to have

pride <<s>> fall

The first finds *expect|expects| expected| expectation* [etc.] preceding *to have*, with up to 6 tokens intervening (+ *expecting, expecte, expectoration*, etc. if they had occurred in that context). The second finds *pride* and *fall* in the same sentence.

[2] http://www.projects.alc.manchester.ac.uk/archer/archer-versions/ and click ARCHER 3.2

Proximity operators include

<<s>>	[= within same sentence
<<n>>	[= within n tokens]
<<n<< and **>>n>>**	[= within n tokens to L or to R]

CQP syntax

For details of the more elaborate CQP ["Corpus Query Processor"] syntax, there is an intro at http://cwb.sourceforge.net/files/CQP_Tutorial/

Here is an easy example:

<s> "(T|t)ake"

The <s> means sense-unit start, so this search finds all sentences beginning with *Take* or *take*.

To create another example, start with a simple query (i.e. not CQP syntax) for the string

could not have

You should get 87 hits. Use your browser's Back button to return to the start page, go to your **Query history**, click on **Show in CQP syntax** and click on the latest query. You are now returned to the search window with the rewritten query ready to run as a CQP query:

[word="could"%c] [word="not"%c] [word="have"%c]

Click **Start query** and get the same hits, as expected. Now go Back and modify the 2nd term only, using **!=** for 'not equals', to search for strings with a different word between the verbs:

[word="could"%c] [word!="not"%c] [word="have"%c]

This gives 59 matches where the word between *could* and *have* is something other than *not*.

Regular expressions

'Regular expressions' or regex have a standard syntax and notation and are widely used in all sorts of programs. Searches in CQPweb (see Hoffmann et al. 2008: 109-13) can make use of regex to capture a complex condition in a single formula.

One of many sites on regex, with brief explanations or full tutorials, can be found at

http://www.regular-expressions.info/

Regex terms are put within square brackets in CQP syntax, **immediately** followed if necessary by one of the quantifiers

?	[= optional]
*****	[= zero or more repetitions]
+	[= 1 or more repetitions]
{n,m}	[= between *n* and *m* repetitions]

Notice how **? * +** behave differently when post-modifying a regular expression than when standing alone as wildcards.

Order and number of hits; distribution by speaker, text type, etc.

Search results are ordered alphabetically by text ID. You can always **sort** the results in various ways that take account of the search string you used and/or its immediate context.

Sometimes, however, you just want to get a feel for a large set of results. Then it's better to use the **Show in random order** button at the top of the concordance (Hoffmann et al. 2008: 53) to avoid getting a skewed impression from the texts (with ARCHER, always the earliest ones) that come first.

If you want to analyse a reasonable number of sentences in detail, then you may have to **thin** a large set of results to get a random selection. Having studied your sample sentences and figured out the proportions of various types within it, you can then extrapolate back to the full set, taking care to be statistically cautious in how much weight you attach to the estimated numbers; see Hoffmann et al. (2008: 80-90) on 'confidence intervals'. You might want to use the Corpus Frequency Wizard at

http://sigil.collocations.de/wizard.html

After any search you can click **Distribution**, then **Go!**, then the appropriate category, for info on speaker age, gender, and many other distributional facts about your hits.

Saving and exporting

Every query you run is saved in your **Query history** (left sidebar on opening page). You can therefore easily re-run a previous query, with or without modifications, and you can also convert a simple query automatically into CQP syntax.

You can also **Save current set of hits** for future use, including the display type and sort in use at the time. Only do this if a lot of work is invested, otherwise it's more economical of your disk space allocation – and almost as quick – just to re-run the query.

To export the results of a query for use in a database program or in MS Word, click **Download | Go!** , fill in the choices (Windows or Mac, etc.), and then click **Download!** The exported text file (extension **.txt** in Windows) can be read by Notepad or Word or other programs.

References

ARCHER website: http://www.projects.alc.manchester.ac.uk/archer/

Hardie, Andrew. 2013. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17.3, 380-409.

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee & Ylva Berglund Prytz. 2008. *Corpus linguistics with BNCweb - a practical guide* (English Corpus Linguistics 6). Frankfurt am Main: Peter Lang.
This book is based on the software BNCweb, which is closely related to CQPweb.