The Corpus Explorer

Lars Nygaard

Version 0.3

About this document

Corpus Explorer (CE) is still under development, and until the interface is more mature, this documentation will remain somewhat scanty. In particluar, it will be somewhat hard to read, since there are not yet any examples.

1 Introduction

CE is a web-based user interface for querying linguistic corpora. Technically, it is a front end for the CQP program, part of the IMS Corpus Workbench[1], a relational database (MySQL[2]) and Ted Pedersen's Ngrams Statistics Package[3].

The development aims have been to create a user interface that is both user friendly and flexible. These two goals are not, however, entirely compatible, however, and the resulting compromise does not allow the user the full range of expression in the CQP search language. Also, the interface is not entirely self-explanatory, so all advanced users should read this manual.

1.1 System requirements

Most modern web browsers can be used to access CE:

- Internet Explorer
- Mozilla, Firefox, Galeon and the rest of the Mozilla family of browsers
- Opera
- Safari

There are currently two exeptions:

- Konqueror
- Internet Explorer for the Macintosh

Konqueror should be supported in future versions; but Internet Explorer for Mac semms to have fallen out of use and will not be supported.

[1] `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench`
[2] `http://mysql.com`
[3] `http://www.d.umn.edu/~tpederse/nsp.html`

2 General options

On the left side of the screen we find options for the entire search. They can be hidden by clicking the small arrow on top of the option field, to create more screen space for building the actual query.

2.1 Source and target corpora

SOURCE CORPUS must be selected. This is the collection of texts that the query expression is first applied to. It can only consist of texts in one language. The queries specified for the TARGET CORPUS are applied to any regions that are aligned to the hits in the source corpus. Several target corpora may be given.

Source and target corpora can be restricted according two criteria:

- language
- original texts or translated texts

2.2 Relation between target query phrases

If more than one target query phrases (see Section 3) are specified for the same target corpus, we can set the relation between them:

conjunction all of the target query phrases must occur in the target corpora

disjunction at least one of the target query phrases must occur in the target corpora.

Source query phrases can only be disjuncted.[4]

2.3 Using regular expression

If the regular expressions box is checked, user input will be interpreted as regular expressions: i.e. "." will be interpreted as "one arbitrary character".

If it is unchecked, all regular expression characters will be *escaped*: i.e. "." will be interpreted as a period.

The regular expression vocabulary is described in Section A.

Queries can sometimes be created faster by typing regular expressions than by selecting items from the menu (eg. typing "house.*" instead of typing "house" and then selecting "start of word" from the menu; typing "house|building" instead of using two query rows).

2.4 Only aligned

Checking this box ("Ignore hits without aligned regions") will restrict the results to hits that have aligned regions in texts that accord with the selection of target corpora.

[4]This is a result of limitations in CQP.

2.5 Search within

This parameter restricts the matching of searches containing arbitrary tokens. It can be set to:

's' where all matches will be within the same s-unit

integer where all matches will be within the specified number of tokens

2.6 Context size

There are two ways of specifying contexts size shown in the search results: by number of sentences or by number of tokens. If we select `s-units`, the `left` and `right` boxes specify the number of sentences to the left and right of the matching sentence.

Similarly, if we select tokens, we specify the number of tokens to the left and right to the matching phrase. Also, the results are displayed as a KWIC concordance.

Context size for target corpus cannot be set; it is always the region or regions aligned to the matching sentence.

2.7 Number of results displayed per page

The search results are divided into a number of pages; the number of results on each page can be adjusted here.

Users of the Mozilla family of browsers[5] should not set this value higher than 20-30 - this will cause the program to display pages quite slowly.

2.8 Maximum number of results

Searching for very common words can be slow. Restricting the total number of results can improve response times.

3 Phrase options

Each "row" in the query builder is a QUERY PHRASE, consisting of several QUERY TOKENS.

Query phrases can be displayed and hidden by clicking the vertical arrows.

Phrases can be specified to apply to the source corpus or the target corpus (except the first phrase, which must apply to the source corpus). If more than one phrase is specified for the source corpus, at least one of the phrases must result in a hit. They are connected by a disjunction operator, to become the SOURCE QUERY.

The target corpus phrases are combined to form the TARGET QUERY. If queries are specified for different parts of the target corpus, all queries must mach (they are connected by the conjunction operator); queries for the same target corpus, or the target corpus as a whole, can be either disjuncted or conjuncted (see 2.2).

[5] Netscape, Firefox, Mozilla, Galeon etc.

Target phrases can be specified to be negative or positive, while source phrases can only be positive.[6]

Additionally, target phrases can be specified to HIGHLIGHT. The phrase will be ignored in the actual corpus search, but matches will be shown in red on the results page.

4 Token options

In addition to simply typing a value in the search field, users can restrict the search further by clicking the "options" box, and selecting values from the menu. Most options can optionally be negated; in this case they will appear with a prefixed exclamation mark ('!').

Options can be removed by double-clicking them.

More token boxes can be added or removed by clicking the horizontal arrows on the right side of the screen.

4.1 Search string options

Start of word if "cat" is entered in the search string box and "start of word" is selected, the program will also return "cats", "category" etc.

End of word if "cat" is entered in the search string box and "end of word" is selected, the program will also return "housecat", "muscat" etc.

Case sensitive if "cat" is entered in the search string box and "case sensitive" is selected, the program will return "cat", but not "Cat".

Lemma if "cat" is entered in the search string box and "lemma" is selected, the program will return all forms of the word, i.e. "cat", and "cats".

4.2 Annotation options

There are two classes of annotation options:

Part of speech gives access to annotations for

- adjective
- adverb
- article
- conjunction
- determiner
- interjection
- infinitive marker
- noun
- particle
- pronoun

[6] This is a result of limitations in CQP.

- proper noun
- preposition
- subjunction
- verb

Miscellaneous gives access to annotations for

- abbreviation
- foreign word
- non-word
- numeral
- possessive ending
- punctuation
- symbol

Positive annotation options will be connected by disjunction, negative annotation options with discjunction. Thus selecting "noun", "verb", "!adjective", "!adverb" will return words that are either nouns or verbs, but neither adjectives nor adverbs.

4.3 Other options

Occurrences allows specification of how many times the token can occur.

Sentence position allows specification of whether or not the token occurs at the start or end of a sentence.

Additional word allows for the input of more than one word form or lemma.

5 Interval options

The minimum and maximum interval specifies the number of unspecified tokens between two token query tokens.

If both are left empty, it is assumed that no unspecified tokens can come between the query tokens (i.e. max: zero, min: zero).

If the minimum interval is specified, but not the maximum, unlimited maximum interval is assumed. Conversely, if the maximum interval is specified, but not the minimum, a minimum interval of zero is assumed.

6 Entering search queries directly

Since the interface does not allow the user the full range of expression in the CWB search language, a separate interface has been created where search expressions can be entered directly.

7 Browsing results

7.1 The results page

The results page consists of:

- A table of token statistics for the queried corpora
- The CQP search string
- A list of available actions (described in section 8).
- A list of results pages, with the current page shown in bold.

7.2 Additional information about hits

Each hit starts with the sentence id. If this id is clicked, a window appears showing meta-information about the text in which the sentence appears. Additionally, it shows more context (and the user can set the context size to an arbitrary large number).

Lemma and part-of-speech of individual tokens are displayed when the mouse is moved over them.

8 Processing results

8.1 Count

This action generates statistics over the matching phrase in the source corpus.

Statistics can be generated for

- word form
- lemma
- part-of-speech
- any combination of the above

The results can be presented in any of the following data formats:

- HTML
- Tab-separated values
- Comma-separated values
- Excel spreadsheet
- Histogram
- Pie chart

8.2 Download

The entire result set can be downloaded, in any of the following data formats:

- Tab-separated values
- Comma-separated values
- Excel spreadsheet

Optionally, additional meta-data may be included in the downloaded result set.

The text in the result set can include

- word form
- lemma
- part-of-speech
- any combination of the above

8.3 Sort

The sorting function applies to the order of the matches in the results set. The set can be sorted alphabetically, according to the source corpus hits, by

- left context
- right context
- matching phrase
- sentence id

When sorting by context or matching phrase, the sorting can be done according to:

- word form
- lemma
- part-of-speech
- any combination of the above

By default, context sorting is done according to the token that is closest to the matching phrase, but the position in context can be set higher by the user.

If the search criteria of two hits are identical, the secondary search criterion applies, with the same options as the primary criterion.

8.4 Collocations

The collocation function compiles statistics of tokens occurring within a user specified context window of the matching phrase.

The available collocation statistics for bigrams are:

- Frequency (no association measure)
- Dice coefficient
- Fisher's exact test
- Log-likelihood ratio
- Mutual information
- Pointwise mutual information
- Odds ratio
- Phi coefficient
- T-score
- Pearson's chi squared test

The available collocation statistics for trigrams are:

- Frequency (no association measure)
- Log-likelihood ratio

The association measures are described in the Ngram Statistics Package documentation `http://search.cpan.org/dist/Text-NSP/Docs/Measures.pod`.

Note that only first word in the matching phrase is used. Thus if any of the matching phrases contain more than one word, the right-side statistics will contain errors.

Statistics can be compiled for:

- word form
- lemma
- part-of-speech
- any combination of the above

The results can be presented in any of the following data formats:

- HTML
- Tab-separated values
- Comma-separated values
- Excel spreadsheet
- Histogram

8.5 Co-occurrence

The co-occurrence functions provide statistics of the words in the target corpus hits.

The available co-occurrence statistics are:

- Frequency (no association measure)
- Dice coefficient
- Fisher's exact test
- Log-likelihood ratio
- Mutual information
- Pointwise mutual information
- Odds ratio
- Phi coefficient
- T-score
- Pearson's chi squared test

The association measures are described in the Ngram Statistics Package documentation `http://search.cpan.org/dist/Text-NSP/Docs/Measures.pod`.

Statistics can be compiled for:

- word form
- lemma
- part-of-speech
- any combination of the above

A rudimentary filter function is also included, restricting the output of the statistics.

The results can be presented in any of the following data formats:

- HTML
- Tab-separated values
- Comma-separated values
- Excel spreadsheet
- Histogram

8.6 Annotate

The annotate function provides a way of marking the sentences in the source corpus according to a user-specified set of values.

Note that this feature is experimental, and is *only* provided for testing purposes. Also, it should be noted that the annotation sets are public and all changes are available for all users.

Acknowledgements

The development has been partially financed by the SPRIK project `http://www.hf.uio.no/forskningsprosjekter/sprik/`.

A Regular expressions

A full account of the regular expressions used by CQP can be found on the IMS website `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/`.

A.1 Optionality

The period (".") represents any character. Thus `.ats` will match "cats", "mats", "bats" etc.

A list of alternative characters can be represented with square brackets: `[cm]ats` will match either "cats" or "mats".

A list of alternative strings can be represented with the vertical bar: `cats|mats` will again match either "cats" or "mats".

A.2 Occurrences

The number of times characters can occur can be specified with the following operators:

? `cats?` matches both "cat" and "cats"

***** `the*` matches "th", "the", "thee" etc.

+ `the+` matches "the", "thee", etc.

{n,n} `the{1,2}` matches "the" or "thee".

A.3 Escaping operators

All the regular expression operators can be searched for; they are interpreted literally if they are prefixed by a backslash. Thus `\.` maches a period in the corpus, and `\?` matches a question mark.