CorpusWeb

The corpus query component of

The Linguistics Web Portal

WEB CWB PORTAL

Contact:
Dr. Sabine Bartsch
Institut für Sprach- und Literaturwissenschaft

Corpus Web
designed for querying corpora

Version date: 04.01.2010 01:57

CorpusWeb is part of The LinguisticsWebPortal which is comprised of the different components, some of which are under construction:

(1) CorpusWeb
(2) LinguisticsGlossary (the linguistics tools and terminology wiki)
(3) LinguisticsTutorials ((Corpus- und Computational-) Linguistics Training Units)
(4) LinguisticsReferences (Linguistics Bibliography)

CorpusWeb is the corpus query component of The Linguistics Web Portal, a portal designed for teaching and learning corpus linguistics. CorpusWeb provides a platform for making linguistic corpora accessible; it offers a three-level corpus query interface that allows the specification of corpus queries of different levels of complexity. The technological backbone of CorpusWeb is the well-established IMS Corpus WorkBench, now known as The IMS Open CorpusWorkBench and its underlying query processor called CQP (Corpus Query Processor) developed at IMS Stuttgart.

Preliminary steps: Being granted access to CorpusWeb

Access to CorpusWeb is granted on an individual user basis which means that each user has to be individually registered in order to be able to access the system. The main reason for this is that many of the corpora made available via CorpusWeb are copyright protected. By granting access to known users only, it is ensured that the legal rights of the copyright owners are respected. Users wanting to get access to CorpusWeb should contact

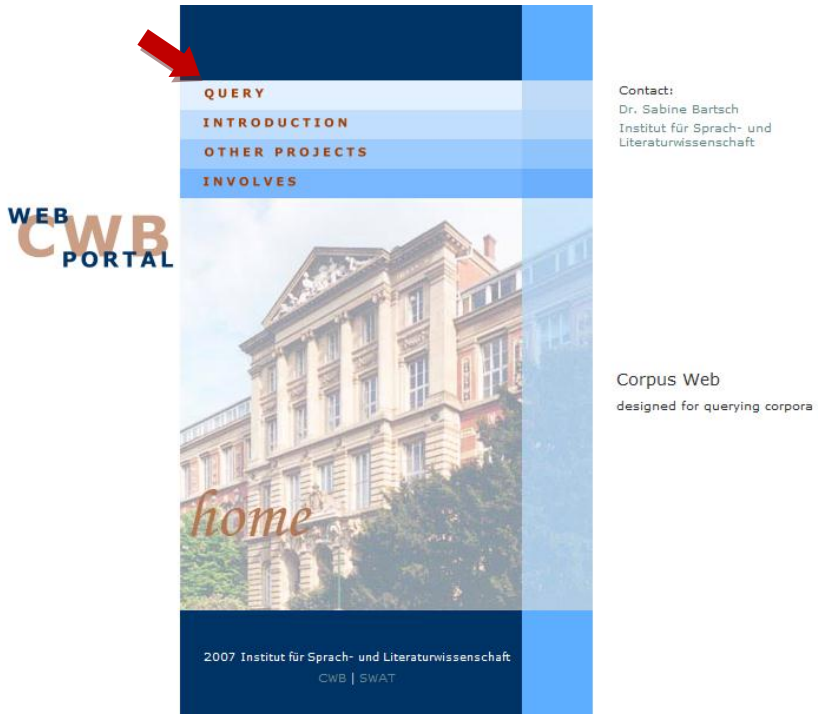
Dr. Sabine Bartsch
Institut für Sprach- und Literaturwissenschaft
Hochschulstrasse 1
64289 Darmstadt
URL: http://www.linglit.tu-darmstadt.de/index.php?id=bartsch

Getting started:

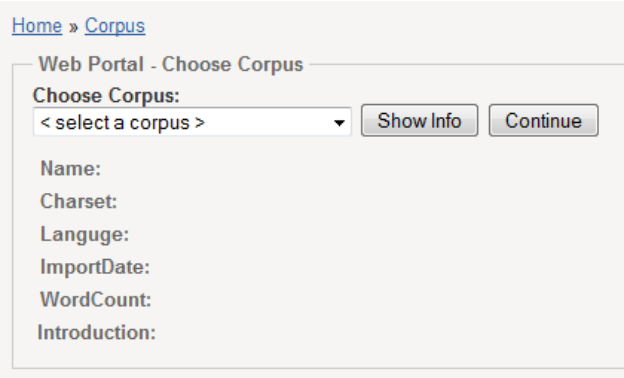
Once an individual username and password are set up, CorpusWeb can be accessed via the Wiki linguisticsweb.org. URL:

http://www.linguisticsweb.org

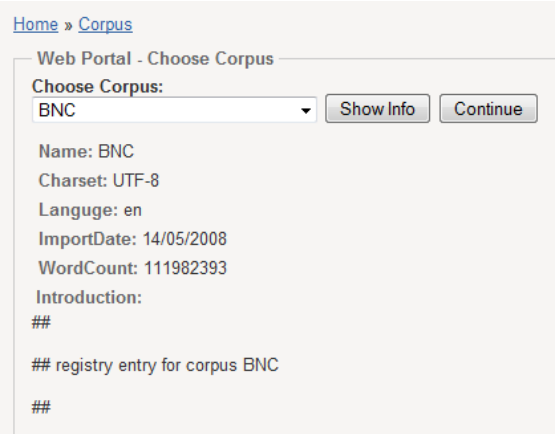
If you are reading this as an online pdf document, click on the URL or copy this line into the address line of your browser, hit ENTER and you should be seeing the following page:



Click on the first menu item QUERY to move on to the corpus selection page:



Click on the drop-down menu text <select a corpus>, the menu will open and show a list of corpora available. Select a corpus, e.g. the BNC (British National Corpus) and click on the button Show Info to get some basic information on the corpus selected. This currently does not look very pretty, but serves the purpose of informing the user of the basic properties of the corpus such as its character encoding and the overall number of words or items found in the corpus:



Once you have selected a corpus, click on Continue. This will get you to the actual corpus query area, the heart of CorpusWeb.

Note that you will find three levels of query tabs at the top of this page:

Simple Query

Advanced Query

Customized Query

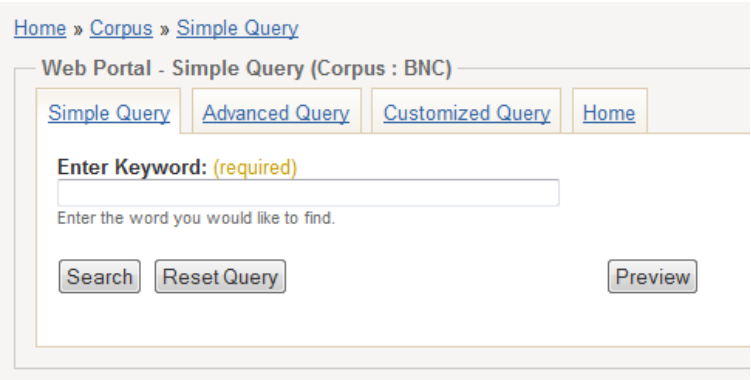
These tabs offer query options at increasing levels of complexity. Simple Query allows you to run simple string queries as well as regular expression queries on just the lexical items in the corpus. It is the starting point into corpus querying. Advanced query allows you to query the corpus for strings / words as well as annotations such as part of speech tagging as well as structures in terms of word or part of speech sequences etc.

Both Simple and Advanced Query allow you to get a preview of the actual query syntax underlying the query you are posing with the guidance of the interface by clicking on the Preview button. The Preview button allows you a view of the query syntax of the query processor that is the backbone of CorpusWeb. This is called CQP (Corpus Query Processor) which is part of the IMS Corpus Workbench. One of the aims of CorpusWeb is to acquaint users with CQP syntax and eventually enable them write more complex queries in the Customized Query box or the actual command line.

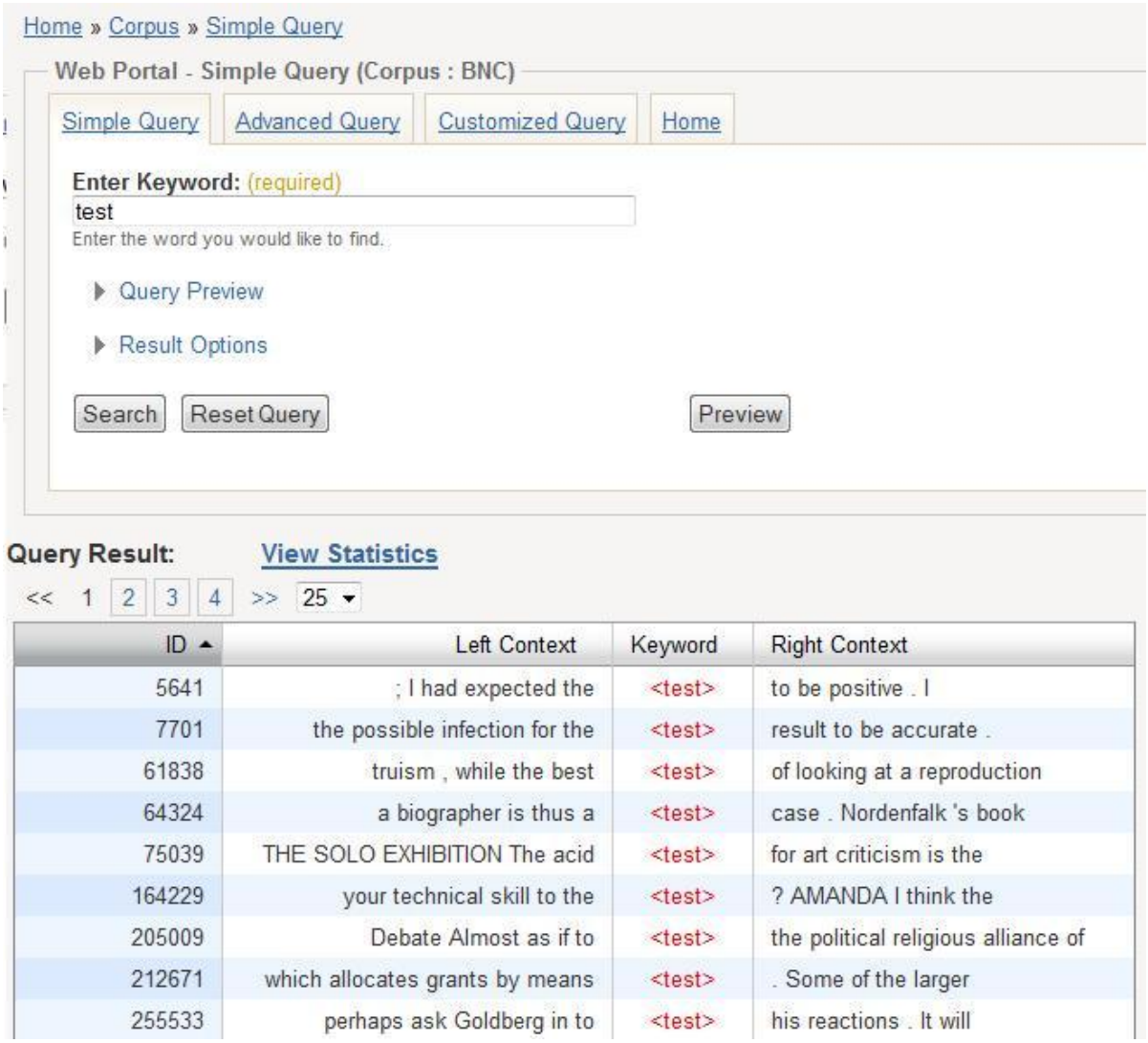
Customized query constitutes a query box that requires queries to be entered in CQP syntax. This allows you to formulate more fine grained and complex queries.

Simple Query

Simple query allows you to query the corpus for simple words. This runs a basic string query which means that it will look for the exact sequence of letters you enter.



You are ready to submit your first query now. Just type a word into the box under **Enter Keyword** and hit the Search button. Depending on the size of the corpus selected and the frequency of the search word in the corpus, the query may take a second or two. When the query completes you can view a concordance view of all occurrences of your query in the lower half of the screen:



ID	Left Context	Keyword	Right Context
5641	; I had expected the	<test>	to be positive . I
7701	the possible infection for the	<test>	result to be accurate .
61838	truism , while the best	<test>	of looking at a reproduction
64324	a biographer is thus a	<test>	case . Nordenfalk 's book
75039	THE SOLO EXHIBITION The acid	<test>	for art criticism is the
164229	your technical skill to the	<test>	? AMANDA I think the
205009	Debate Almost as if to	<test>	the political religious alliance of
212671	which allocates grants by means	<test>	. Some of the larger
255533	perhaps ask Goldberg in to	<test>	his reactions . It will

Play around with different queries. In order to enter a new query, it is a good idea to regularly hit the Reset Query button before a new query in order to clear the previous query from the cache.

Query Result:

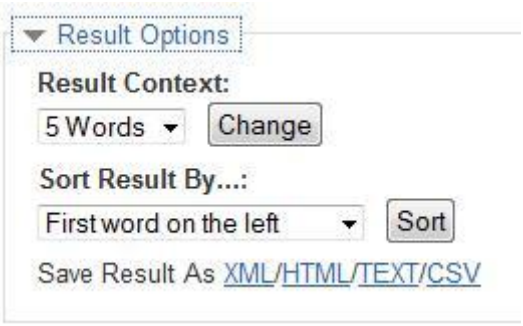
The result of your query is presented in classic key-word in context (KWIC) concordance format at the bottom of the tab. By default, 25 result lines are shown per window. You can, however, change this setting to up to 100 lines per window by clicking on the respective drop down field above the concordance under Query Result. In this area, you can also move from one result window to the next to view all results by clicking on the consecutive page numbers or the arrows to move back and forth.

You can also sort according to all four result columns by clicking on one of the headers of the result table. As an example, click on the box Right Context and CorpusWeb will sort your results alphabetically according to the sequence of characters in the right context. The same action can be carried out for the Left Context and for the Keyword, although the latter function makes more sense when your query yields different forms of the keyword or different keyword, e.g. all nouns in the corpus. You can also sort according to the left-most column entitled ID. This function builds on the

underlying index of the corpus and allows you to view all results in ascending or descending corpus order. This can be very useful if you need to retrace the origin of your examples.

The option View Statistics is currently under construction as it does not work with all versions of Adobe Flash Player which is required for a graphical presentation of the quantitative results.

Further options for the presentation of the results are available via field of functions entitled **Result Options** further up on the tab:

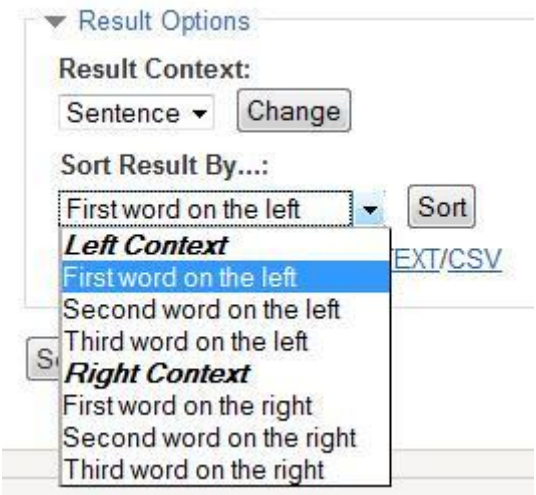


This toggles three types of setting:

Result Context allows you to alter the size of the context shown in the concordance to the left and right of the keyword. By default it is set to 5 Words, but the drop-down list allows you the set this to 10 Words or the full Sentence context.

Sort Result By …:

This function allows you to sort more specifically according a given word in the left or right context. The drop-down menu allows you to set the sorting algorithm to the first, second or third word to the left and right context.



Save Result As XML / HTML / TEXT / CSV:

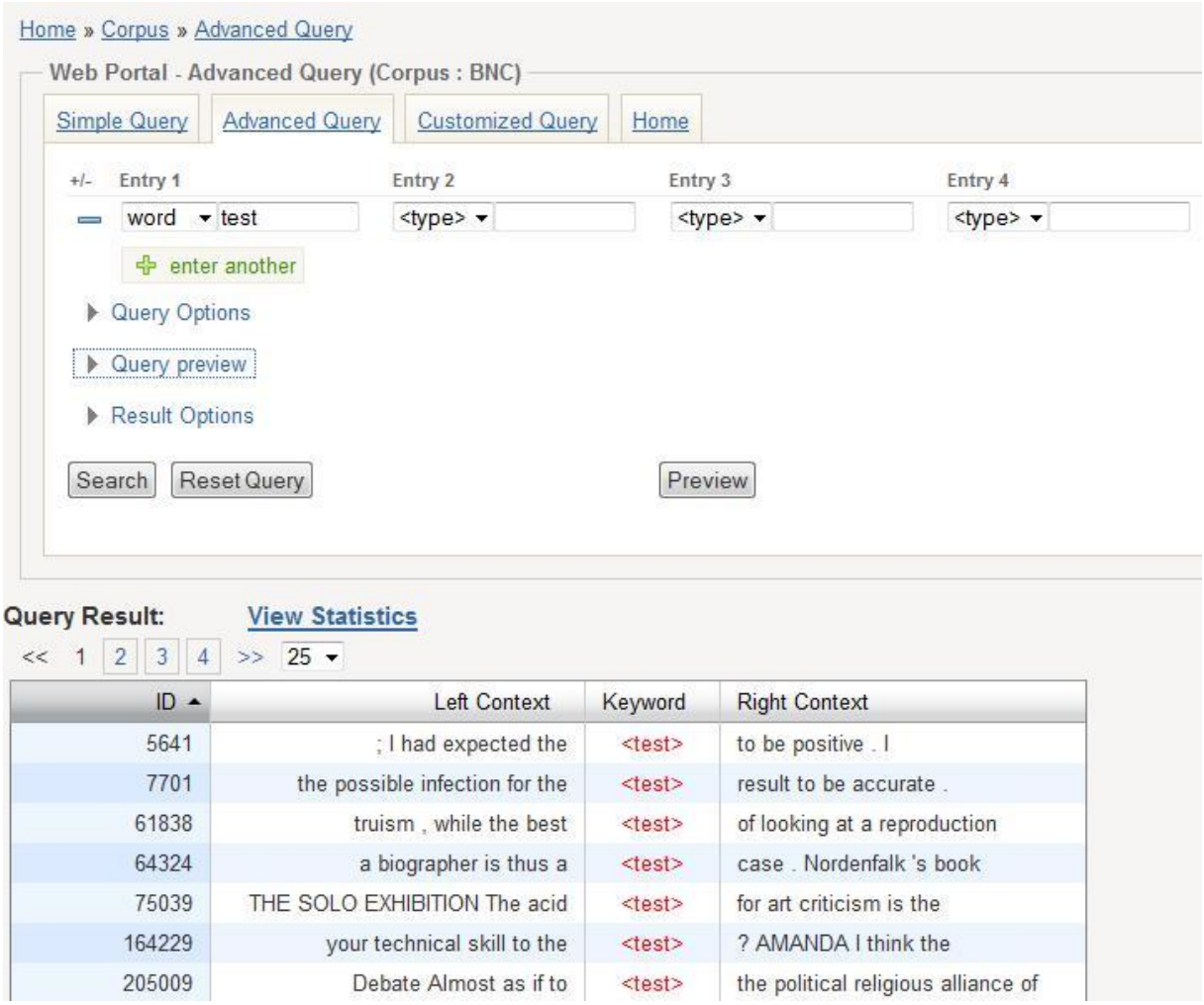
The third function allows you to save the results in the concordance in a choice of standard formats.

Another function available via a button in this tab is the Query Preview:

Query Preview allows you to view the syntax or your query as it is posed by the system. This function was implemented in order to aid learners in familiarizing themselves with the syntax of the query system underlying CorpusWeb called CQP (Corpus Query Processor). It is explained in more detail further down in this Quick Start Guides.

Advanced Query

The Advanced Query tab offers more complex query options for words in context and for the annotations in an annotated corpus.

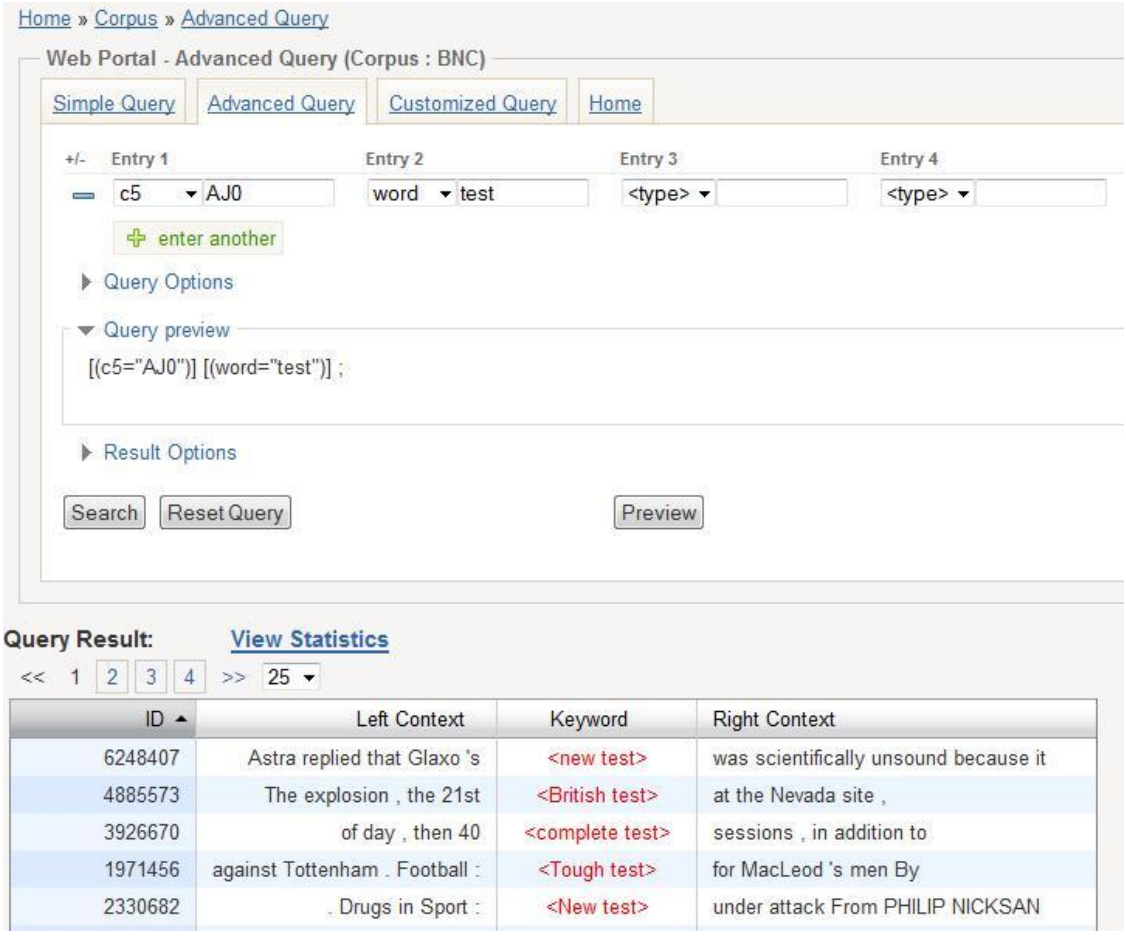


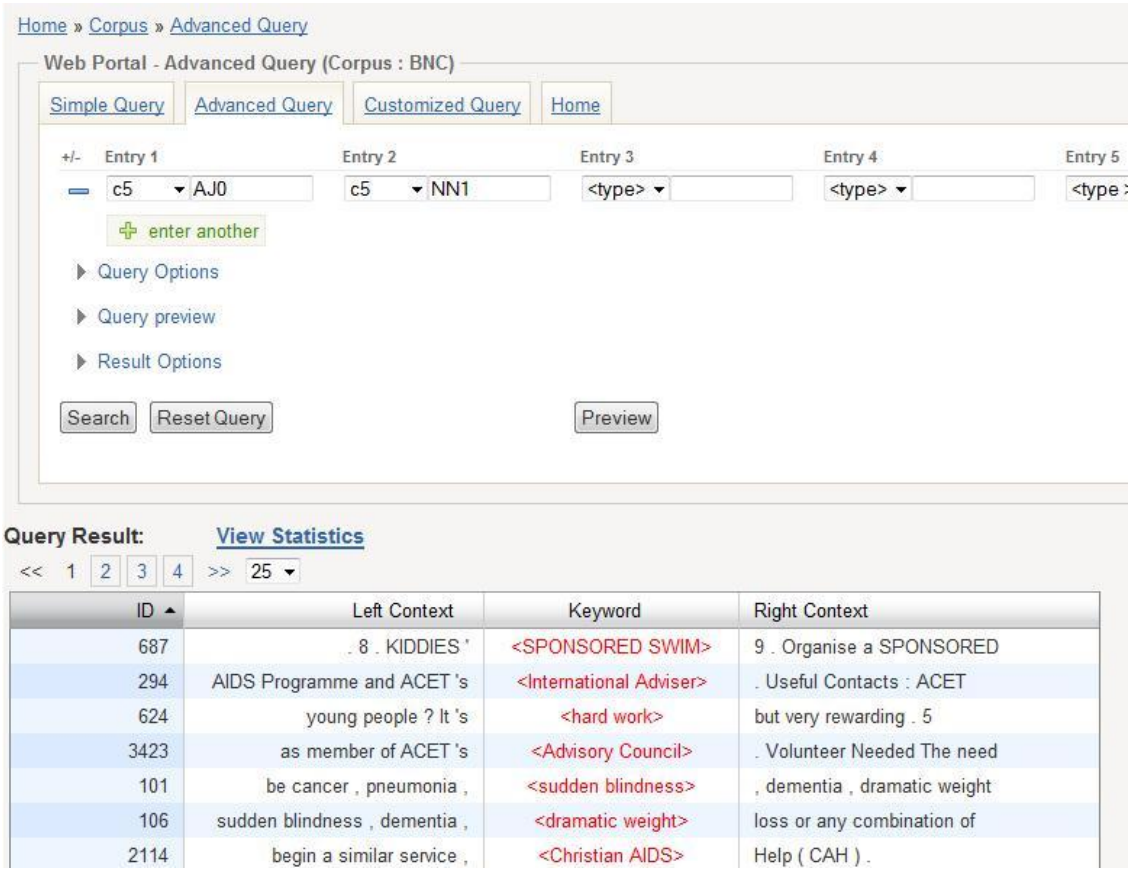
In the simplest case, you just query for a word by selecting word from the left drop-down field and enter the word you are looking for in the box to the right of the the first entry **Entry 1**. In the case illustrated in the figure above, we are querying for just the word *test*.

Each entry, **Entry 1** – **Entry 5** in sequence thus stands for one word in a sequence of words which can be identified either as a string or by means of features attributed to it as a type of annotation. In case of the British National Corpus, we can query for the following features:

Attribute	Definition
word	String or regular expression, e.g. test which find just instances of the word *test* the regular expression *test.** which finds all strings starting with test and ending in any sequence up to the next white space (e.g. *test*, *tests*, *testing*, *tested*, *tester*, *testbed*, *testcase* etc.)
c5	Any tag according to Claws 5, the annotation scheme used by the CLAWS tagger developed at Lancaster University and employed in the annotation of the British National Corpus. An overview of the C5 Tagset used in the British National Corpus can be found at the following URL: http://www.natcorp.ox.ac.uk/docs/bnc2guide.htm#tagset
lemma	The lemma, i.e. the canonical or base for of a word. Entering the lemma test finds all forms of the word: *test*, *tests*, *testing*, *tested*, but not its compound forms *testbed*, *testcase*, because these are different words which have their own lemma forms.
pos	In the case of the BNC, this a simplified part of speech annotation in the verbatim form Verb, Noun, Adjective etc.
sid	This builds again on the index underlying CQP. Each item in a corpus has a unique ID to which you can trace it back, e.g. if you want to retrace the origin of a word in a specific sentence.

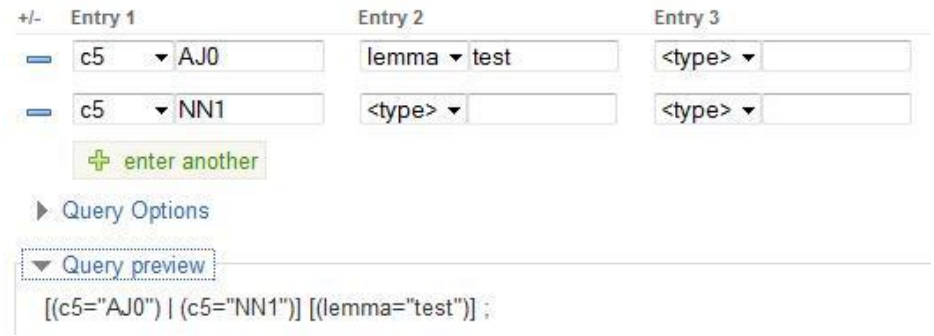
You can also query for combinations of different features, e.g. a specific part of speech preceding a specific word. In the example illustrated in the image below, we are querying for any adjective preceding the word *test*. We are thus querying a combination of words with a specific part of speech tag followed syntagmatically by a specific word, in this case the word *test*. This can be very useful if you want to find out with which adjectives a specific noun tends to collocate.





Advanced Query also allows you to combine different queries into one. For example, you might want to query simultaneously for the lemma test (lemma = “test”) preceded by either an adjective in base form (c5 = “AJ0”) or a noun in the singular (c5 = “NN1”). All you have to do is click on the button 

to open up a new line of Entries. This will open up a new line of entry boxes into which you can enter a word on the position you want to query. In this case, you are querying for two types of Entry 1 followed by the same lemma in Entry 2:



The Query Preview shows you the query syntax. Under the bonnet, the query looks like this:

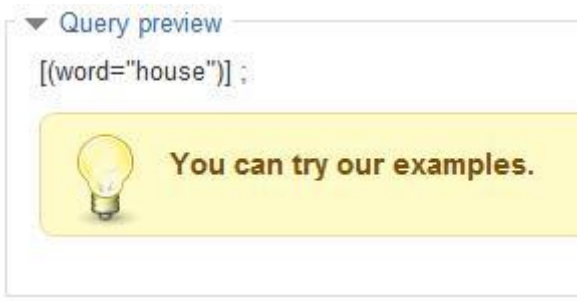
[(c5="AJ0") | (c5="NN1")] [(lemma="test")] ;

The query for either AJ0 or NN1 is indicated by the pipe symbol (|) between the two query options in Entry 1, the two queries are combined by means of entering them between square brackets []; the query for the following lemma of the word test is expressed in the second set or square brackets.

This is an example of the syntax underlying the whole system. It is the kind of syntax that must be used in the next tab Customized Query described below.

Customized Query

As was already mentioned above, the system in the background of this interface is CQP, the Corpus Query Processor, a very powerful query engine that allows you to formulate much more advanced queries than the one you are currently experimenting with. It has a complex query syntax which the CorpusWeb interface writes for you in the background. If you would like to see what your query looks like in real CQP syntax, you can always click on the Preview button which brings up a window that shows the full CQP version of your query:



To get started with the functionality of the Customized Query tab, please type a query into the box exactly as you see it in the Preview on the Advanced Query tab.

Example$_1$: [(word=”house”)] ;

Example$_2$: [(c5="AJ0") | (c5="NN1")] [(lemma="test")] ;

Bibliography and sources:

CLAWS5 Tagset used in the British National Corpus: URL: http://www.natcorp.ox.ac.uk/docs/bnc2guide.htm#tagset

IMS Corpus WorkBench (CWB). URL: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

Oli Christ: *"A modular and flexible architecture for an integrated corpus query system". COMPLEX'94, Budapest, 1994.*

Oli Christ and B.M.Schulze: *"Ein flexibles und modulares Anfragesystem für Textcorpora"*. Tagungsbericht des Arbeitstreffen Lexikon + Text. Niemeyer, Tübingen, 1995.

IMS Open CorpusWorkBench. URL: http://cwb.sourceforge.net/

CQP Query Syntax:

http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPSyntax.html

CQP Example Queries:

http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPExamples.html