

Corpus Linguistics: A Practical Introduction

Nadja Nesselhauf, October 2005 (last updated September 2011)

1) Corpus Linguistics and Corpora

- What is corpus linguistics (I)?
- What data do linguists use to investigate linguistic phenomena?
- What is a corpus?
- What is corpus linguistics (II)?
- What corpora are there?
- What corpora are available to students of English at the University of Heidelberg?

(For a list of corpora available at the Department of English click [here](#))

2) Corpus Software

- What software is there to perform linguistic analyses on the basis of corpora?
- What can the software do?
- A brief introduction to an online search facility (BNC)
- A step-to-step introduction to WordSmith Tools

3) Exercises (I and II)

- I Using the WordList function of WordSmith
- II Using the Concord function of WordSmith

4) How to conduct linguistic analyses on the basis of corpora: two examples

- Example 1: Australian English vocabulary
- Example 2: Present perfect and simple past in British and American English
- What you have to take into account when performing a corpuslinguistic analysis

5) Exercises (III)

- Exercise III.1
- Exercise III.2

6) Where to find further information on corpus linguistics

1) Corpus Linguistics and Corpora

What is corpus linguistics (I)?

Corpus linguistics is a method of carrying out linguistic analyses. As it can be used for the investigation of many kinds of linguistic questions and as it has been shown to have the potential to yield highly interesting, fundamental, and often surprising new insights about language, it has become one of the most wide-spread methods of linguistic investigation in recent years.

What data do linguists use to investigate linguistic phenomena?

Roughly, four types of data for linguistic analysis can be distinguished:

- 1) data gained by intuition
 - a) the researcher's own intuition ("introspection")
 - b) other people's ("informant's") intuition (accessed, for example, by elicitation tests)
- 2) naturally occurring language
 - a) randomly collected texts or occurrences ("anecdotal evidence")
 - b) systematic collections of texts ("corpora")

(For further reading on corpora vs. intuition, see Fillmore 1992)

What is a corpus?

A corpus can be defined as a systematic collection of naturally occurring texts (of both written and spoken language).

"Systematic" means that the structure and contents of the corpus follows certain extralinguistic principles ("sampling principles", i.e. principles on the basis of which the texts included were chosen). For example, a corpus is often restricted to certain text types, to one or several varieties of English, and to a certain time span. If several subcategories (e.g. several text types, varieties etc.) are represented in a corpus, these are often represented by the same amount of text. "Systematic" also means that information on the exact composition of the corpus is available to the researcher (including the number of words in each category and in the whole corpus, how the texts included in the corpus were sampled etc).

Although "corpus" can refer to any systematic text collection, it is commonly used in a narrower sense today, and is often only used to refer to systematic text collections that have been computerized.

What is corpus linguistics (II)?

Corpus linguistics thus is the analysis of naturally occurring language on the basis of computerized corpora. Usually, the analysis is performed with the help of the computer, i.e. with specialised software, and takes into account the frequency of the phenomena investigated.

What corpora are there?

There are many types of corpora, which can be used for different kinds of analyses (cf. Kennedy 1998). Some (not necessarily mutually exclusive) examples of corpus types are (for a description of the individual corpora see below):

- general/reference corpora (vs. specialized corpora)
(e.g. BNC = British National Corpus, or Bank of English):
 - aim at representing a language or variety as a whole (contain both spoken and written language, different text types etc.)
- historical corpora (vs. corpora of present-day language)
(e.g. Helsinki Corpus, ARCHER)

- aim at representing an earlier stage or earlier stages of a language
- regional corpora (vs. corpora containing more than one variety)
(e.g. WCNZE = Wellington Corpus of Written New Zealand English)
aim at representing one regional variety of a language
- learner corpora (vs. native speaker corpora)
(e.g. ICLE = International Corpus of Learner English)
aim at representing the language as produced by learners of this language
- multilingual corpora (vs. one-language corpora)
aim at representing several, at least two, different languages, often with the same text types (for contrastive analyses)
- spoken (vs. written vs. mixed corpora)
(e.g. LLC = London-Lund Corpus of Spoken English)
aim at representing spoken language

A further distinction of corpus types refers not to the texts that have been included in the corpus, but to the way in which these texts have been treated:

- annotated corpora (vs. orthographic corpora)
in annotated corpora, some kind of linguistic analysis has already been performed on the texts, such as sentence analysis, or, more commonly, word class classification

What corpora are available to students of English at the University of Heidelberg?

1) Generally accessible corpora:

Two large general corpora of English are accessible to everyone via the World Wide Web. These are the Collins Wordbanks Online English corpus and the British National Corpus.

- a) The Collins Wordbanks Online English corpus contains 56 million words of contemporary written and spoken text, both British and American English, of a variety of text types. Of these, 36 million are British written texts, 10 million American written texts and 10 million American spoken texts. The user can select either one or two or all three subcorpora for the analysis. The corpus is accessible at: <http://www.collins.co.uk/Corpus/CorpusSearch.aspx>.
- b) The British National Corpus (BNC) contains 100 million words of contemporary British English, of which 90 million are written and 10 million spoken texts (of a variety of different text types). For simple searches, the corpus is accessible at: <http://sara.natcorp.ox.ac.uk/lookup.html> (For an explanation of the search facility, see below. Much more complex searches are possible on the basis of the CD-Rom version of the corpus, also available at the Department).

A few smaller and more specialised corpora have also been made available on the internet. An example is MICASE:

- c) MICASE (The Michigan Corpus of Academic Spoken English) contains 1.8 million words of (transcribed) academic speech, recorded at the University of Michigan between 1997 and 2001. The corpus contains different academic speech events, and both native and non-native speaker language. It is available at: <http://micase.umdl.umich.edu/m/micase/>.

2) Corpora at the English Department

A great number of corpora are available for students of English, the most important of which are listed below. Many of these are accessible on one of the two “corpus computers” located in the seminar library (“library” in the table).

The corpus computers are located in the second room, in the right-hand corner, just before the linguistic section of the library. Other corpora are available at the Lehrstuhl Prof. Busse.

If you intend to carry out research with one of the corpora, please contact Nadja Nesselhauf:

Nadja.Nesselhauf@urz.uni-heidelberg.de

For office hours, see: <http://www.as.uni-heidelberg.de/personen/>

You will then receive a password for the corpus computers or receive access to one of the corpora available at the Lehrstuhl.

List of corpora available at the Department of English

Most of the available corpora are stored directly in C: on the corpus computers (the path for ICE New Zealand, for example, simply is C:\icenz), many are stored in C:\ICAME\texts. Many manuals are also stored on the corpus computer, in the same directory as the corpus in question.

Extensive lists of existing corpora can be found on Michael Barlow’s “Corpus Linguistics Site” at <http://www.athel.com/corpus.html>) and on the “English language corpora and corpus resources” page at <http://www.natcorp.ox.ac.uk/corpora.html>.

Corpus	Variety/ies	Time (span)	Number of words	Text type(s)	Available	Further information available at / in:
Large general corpora:						
BNC (British National Corpus)	BrE	PDE	100 m	90% written, 10% spoken, many diff. text types	General access	simple searches: http://www.natcorp.ox.ac.uk/using/index.xml?ID=simple more advanced searches: http://bncweb.lancs.ac.uk/bncwebSignup/user/login.php
COCA (The Corpus of Contemporary American English)	AmE	1990-2011	425 m	Various spoken and written text types	General access	http://www.americancorpus.org/
ANC (American National Corpus)	AmE	PDE	20 m	Written and spoken	librarian	http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T20
Comparable corpora of written British and American English:						
LOB	BrE	1961	1 m	2,000-word samples from diff. written text types	Corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm manual in library next to corpus computers
Brown	AmE	1961	1 m	same as LOB	Corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm manual in library next to corpus computers
FLOB	BrE	1991	1 m	same as LOB	Corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm manual in library next to corpus computers
Frown	AmE	1992	1 m	same as LOB	Corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm manual in library next to corpus computers
Corpora of other varieties, comparable to the above in composition:						
WWC (Wellington Corpus of Written New Zealand English)	NZE	1986-1990	1 m	same as LOB	corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm manual in library next to corpus computers

Kolhapur	IndE	1978	1 m	same as LOB	corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm
ACE (The Australian Corpus of English)	AuE	1986	1 m	same as LOB	corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm manual in library next to corpus computers
Other regional corpora:						
ICE-New Zealand	NZE	1989-1994	1 m	diff. spoken (600,000) and written (400,000) genres	corpus computer (library)	http://www.ucl.ac.uk/english-usage/ice/index.htm
ICE-Singapore	SingE	PDE	1 m	diff. spoken (600,000) and written (400,000) genres	General access	http://ice-corpora.net/ice/index.htm
ICE-India	IndE	PDE	1 m	diff. spoken (600,000) and written (400,000) genres	General access	http://ice-corpora.net/ice/index.htm
ICE-Philippines	PhilE	PDE	1 m	diff. spoken (600,000) and written (400,000) genres	General access	http://ice-corpora.net/ice/index.htm
ICE-GB (Great Britain)	BrE	1990-1993	1 m	diff. spoken (600,000) and written (400,000) genres	librarian	http://www.ucl.ac.uk/english-usage/ice/index.htm
ICE-Canada	CanE	PDE	1 m	diff. spoken (600,000) and written (400,000) genres	General access	http://ice-corpora.net/ice/index.htm
ICE-Hong Kong	HKE	PDE	1 m	diff. spoken (600,000) and written (400,000) genres	General access	http://ice-corpora.net/ice/index.htm
ICE-Ireland	IrE	PDE	1 m	diff. spoken (600,000) and written (400,000) genres	General access	http://ice-corpora.net/ice/index.htm
ICE-East Africa	Kenya n & Tanz. Engl.	1990-1996	1.3 m	400,000 w. written Kenyan English, 400,000 w. written Tanzanian English, 500,000 words spoken (Kenyan & Tanzanian)	Generally available	http://ice-corpora.net/ice/index.htm

WSC (Wellington Corpus of Spoken New Zealand English)	NZE	1988-1994	1 m	2,000 word extracts of formal, semi-formal, and informal speech	Corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm
Spoken corpora:						
COLT (Corpus of London Teenage Language)	BrE	1991	500,000	transcribed spontaneous spoken lg. of London teenagers	Corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm
LLC (London-Lund Corpus of Spoken English)	BrE	1975-1988	500,000	transcribed spoken lg., various text types	Corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm
SEC (Lancaster / IBM Spoken English Corpus)	BrE	1984-1985	53,000	transcribed spoken lg., primarily radio broadcasts	Corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm
Santa Barbara Corpus of Spoken American English	AmE	PDE	300,000	Transcribed spoken interaction, from all over the U.S.	General access	http://www.linguistics.ucsb.edu/research/sbcampus.html
MICASE (The Michigan Corpus of Academic Spoken English)	AmE	1997-2001	1,8 m	Transcribed academic speech (native and non-native)	General access	http://micase.elicorpora.info/ or http://quod.lib.umich.edu/m/micase/
Historical corpora:						
Helsinki Corpus (diachronic part)	BrE	750-1700	1.5 m	different text types	corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm
Helsinki Corpus of Older Scots	Scots	1450-1700	870,000	several text types	corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm
ARCHER (A Representative Corpus of Historical English Registers)	BrE & AmE	1650-1990	1.7 m	several diff. text types, both written and speech-based	Contact Dr. Nesselhauf	http://www.llc.manchester.ac.uk/research/projects/archer/archer3_1/

CEECs (Corpus of Early English Correspondence Sampler)	BrE	1418-1680	450,000	letters	corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm
PCEEC (Parsed Corpus of Early English Correspondence Sampler)	BrE	1418-1680	2.2 m	Letters	General access	http://ota.ahds.ac.uk/
DCPSE	BrE	1950s-1990s	800,000	Various spoken text types	librarian	http://www.ucl.ac.uk/4english-usage/projects/dcpse/
CED	BrE	1560-1760	1.2 m	various speech-based text types	contact Prof. Busse	http://www.helsinki.fi/varieng/CoRD/corpora/CED/basic.html
Lampeter Corpus of Early Modern English Tracts	BrE	1640-1740	1.1 m	tracts from various areas (religion, law, science etc.)	corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm
ICAMET (Innsbruck Computer Archive of Machine-Readable English Texts)	BrE	1100-1688	5.6 m	prose 1100-1500, letters 1386-1688	corpus computer (library)	http://anglistik1.uibk.ac.at/ahp/projects/icamet/icamet1.html
COHA (The Corpus of Historical American English)	AmE	1810-2009	400 m	Various text types	General access	http://corpus.byu.edu/coha
Concer	BrE	19 th c.		Various text types	Contact Prof. Busse	
18 th Century Fiction	BrE	18 th c.				
19 th Century Fiction		19 th c.				
Scientific and Medical Writings in Old and Middle English	BrE	OE and ME		Scientific and medical writing	librarian	

Middle English Medical Texts	BrE	ME	500,000	Medical writing (surgical treatises, remedy books, recipes etc.)	librarian	http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/MEMTindex.html
Early Modern English Medical Texts	BrE	EMO	2 m	Different types of medical writing	librarian	http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/MEMTindex.html
Learner corpus:						
ICLE (International Corpus of Learner English)	learner Engl.	1990s	2.5 m	argumentative essays written by learners with different L1s	corpus computer (library)	Granger, Sylviane; Dagneaux, Estelle, Meunier, Fanny (2002). <i>International Corpus of Learner English. Handbook</i> . (library, next to corpus computer)
Corpus of child language:						
POW (Polytechnic of Wales Corpus)	BrE	1978-1984	65,000	speech of children between 6 and 12 (spontaneous and interviews)	corpus computer (library)	http://khnt.hit.uib.no/icame/manuals/index.htm

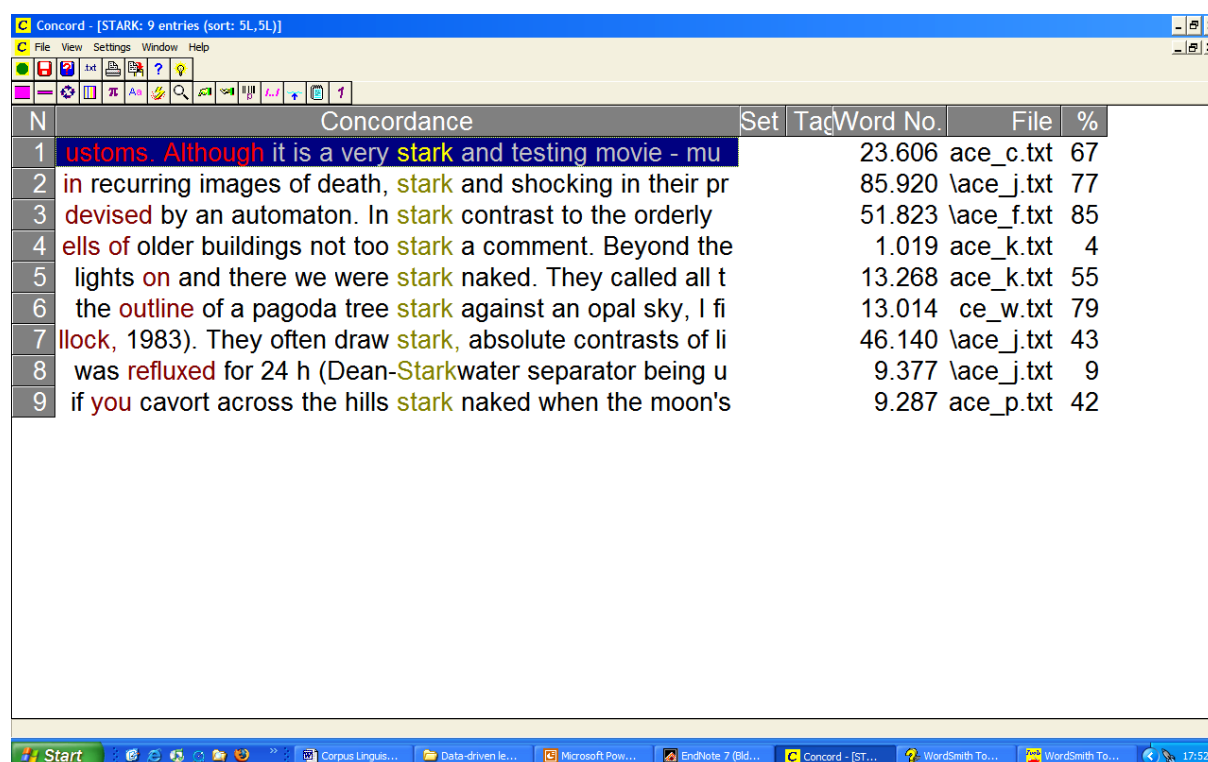
2) Corpus Software

What software is there to perform linguistic analyses on the basis of corpora?

Two types of software for corpus analysis can be distinguished in principle: software that is tailored to one specific corpus, and software that can be used with almost any kind of corpus. Examples of the former type are two software programs that have been tailored to the British National Corpus, namely SARA and BNCWeb (accessible on the left corpus computer in the seminar library of the English Department at the University of Heidelberg). A further example is ICE-CUP, which has been tailored to ICE-GB. A subgroup of this type is software which allows searches in one specific corpus over the internet (such as the online search facilities provided for the BNC and the Collins Wordbanks *Online* English mentioned above). Examples of the second type of more generally usable software are MonoConc Pro (demo available at <http://www.camsoftpartners.co.uk/monoconc.htm>) and WordSmith Tools, which is probably the most widely used corpus software. WordSmith is available for corpus research on the right corpus computer in the library; its use will be explained in detail below.

What can the software do?

While there are many differences between the software packages designed for corpus analysis, certain basic functions can be performed by practically all the available software. For most kinds of linguistic analyses, the most important one of these is the possibility of searching the corpus in question for the occurrence of certain strings (i.e. words or phrases). As output, the software then usually gives information on the number of these strings occurring in the corpus, on the part of the corpus and/or text in which they were found, and so-called concordance-lines, which show the string in question in context (with the search term(s) highlighted), such as in this example (the result of a WordSmith query for the string *stark* in ACE, the Australian Corpus of English):



N	Concordance	Set	Tag	Word No.	File	%
1	ustoms. Although it is a very stark and testing movie - mu	23.606		ace_c.txt	67	
2	in recurring images of death, stark and shocking in their pr	85.920		\ace_j.txt	77	
3	devised by an automaton. In stark contrast to the orderly	51.823		\ace_f.txt	85	
4	ells of older buildings not too stark a comment. Beyond the	1.019		ace_k.txt	4	
5	lights on and there we were stark naked. They called all t	13.268		ace_k.txt	55	
6	the outline of a pagoda tree stark against an opal sky, I fi	13.014		ce_w.txt	79	
7	lock, 1983). They often draw stark, absolute contrasts of li	46.140		\ace_j.txt	43	
8	was refluxed for 24 h (Dean-Starkwater separator being u	9.377		\ace_j.txt	9	
9	if you cavort across the hills stark naked when the moon's	9.287		ace_p.txt	42	

The number on the left indicates the total number of occurrences (9), on the right, further information such as in which file the instances were found is provided (what type of text the

different file names refer to can be found out by consulting the manual; ace_j.txt, for example, means that the instance occurs in a science text, ace_k.txt that it occurs in a fiction text etc.).

Two further basic functions that can be performed by almost all corpus software are sorting (for example according to the word to the right or left of the search term) and “thinning” the results (i.e. the removal of irrelevant instances). In this example, the instance number 8 is irrelevant if the researcher is interested in the adjective *stark*; this line can then be removed from the concordance (see below for a step-to-step introduction to WordSmith). Corpus software can also usually search for words or phrases occurring within a certain distance of each other and also usually allows the use of wildcards in searches, i.e. the use of certain symbols which can represent any one character or word (if “^” is a wildcard representing one letter, a search for *s^ng*, for example, will yield instances of *sing*, *sang*, *sung* and *song*).

The advantage of the analysis of texts with corpus software is apparent: in a few seconds the researcher receives information (about the occurrence of linguistic items in a large amount of text) that would take hours or even days if it had to be retrieved manually. The concordance-line output allows the researcher to see the occurrences in context, so that the use of the linguistic item in question, in particular frequent patterns, can often be investigated with little effort.

Retrieving more context than shown in the concordance lines, searching only a part of a given corpus, and saving the results are also regular features of almost all corpus software. Most programs can also find words frequently occurring in the vicinity of the search term and provide a list of all the words occurring in the corpus and their frequencies.

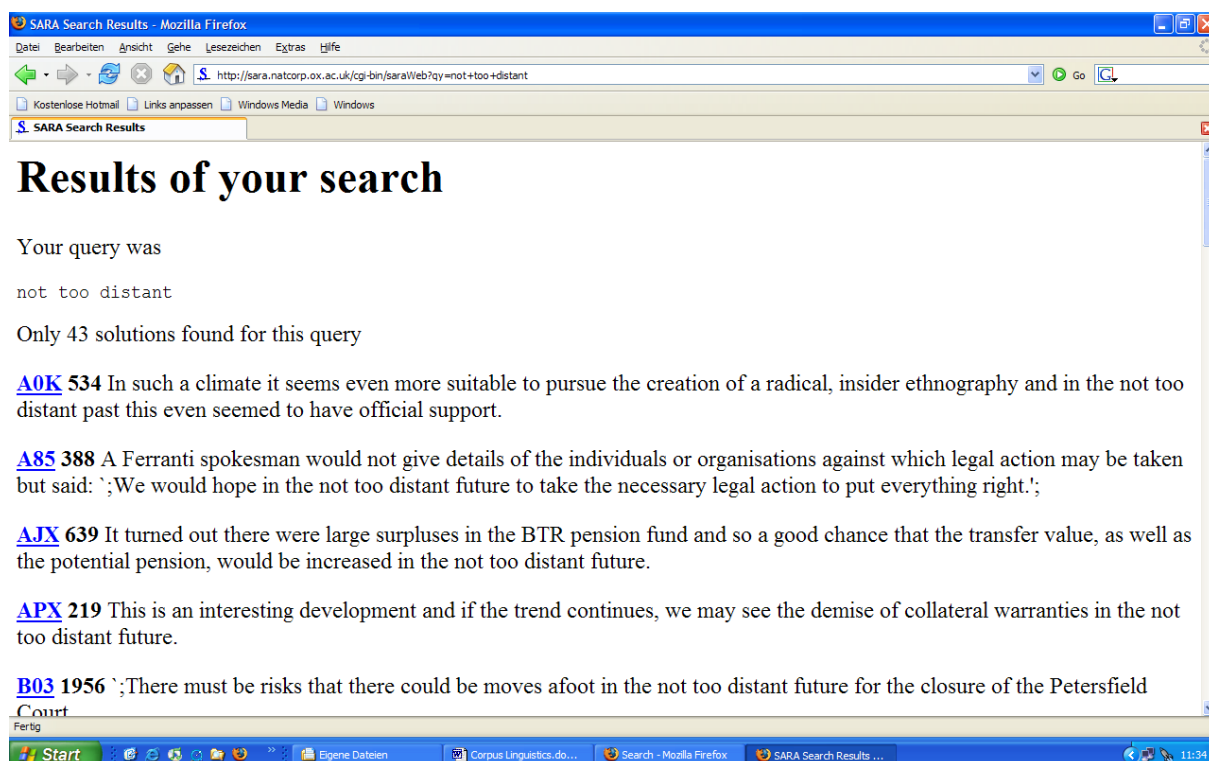
A brief introduction to an online search facility (BNC)

As mentioned above, the BNC (British National Corpus) can be accessed via the world wide web, by an online search facility allowing simple searches (available at <http://sara.natcorp.ox.ac.uk/lookup.html>). The search interface looks as follows:

Simple Search of BNC-World

Please enter your query:

In the space provided, you can enter either words or phrases. A few more complex types of searches, such as for two words with another – unspecified – word intervening, can also be performed (a concise explanation of how to perform such more complex queries is provided on the site). Of the results, 50 at most are displayed; if there are more occurrences in the corpus, the overall number of occurrences is stated, and 50 randomly selected instances are displayed. The results are given in the context of the sentence in which they occur. If you enter the phrase *not too distant*, for example, the results are presented as follows:



If you click on the 3-digit combinations of letters and numbers given on the left, information on the text in which the sentence occurs is given.

The disadvantage of this kind of presentation is that, as the search terms are neither highlighted or aligned, patterns are not easily visible. Unlike the software that can be used with a local version of the BNC, basic functions such as sorting or thinning the results or getting more context cannot be performed, and if there are more than 50 hits not all of them can be inspected. The facility (as well as most other online corpus search facilities available) is therefore not suited for most kinds of more comprehensive linguistic research. It is, however, useful for some first practise in corpus linguistics, and for investigating some kinds of (mostly “small”) linguistic questions such as those which L2 users ask themselves when composing a text (in this case, the question might have been: *Can not too distant* only be used in the phrase *in the not too distant future*, or are there other possible combinations, such as *not too distant past*?). For corpus linguistic studies, this facility as well as others of its kind are useful for a first exploration of whether a certain linguistic feature is worth investigating and which questions as to the use of this feature might yield interesting results.

A step-to-step introduction to WordSmith Tools

1) Getting started

To start WordSmith (for example on the right corpus computer in the library of the English Department):

- either doubleclick on the yellow “WordSmith Tools” Icon
- or choose “Start” – “Programme” – “WordSmith Tools”

Several windows will pop up.

- minimize the first window (“WordSmith Tools Help”)

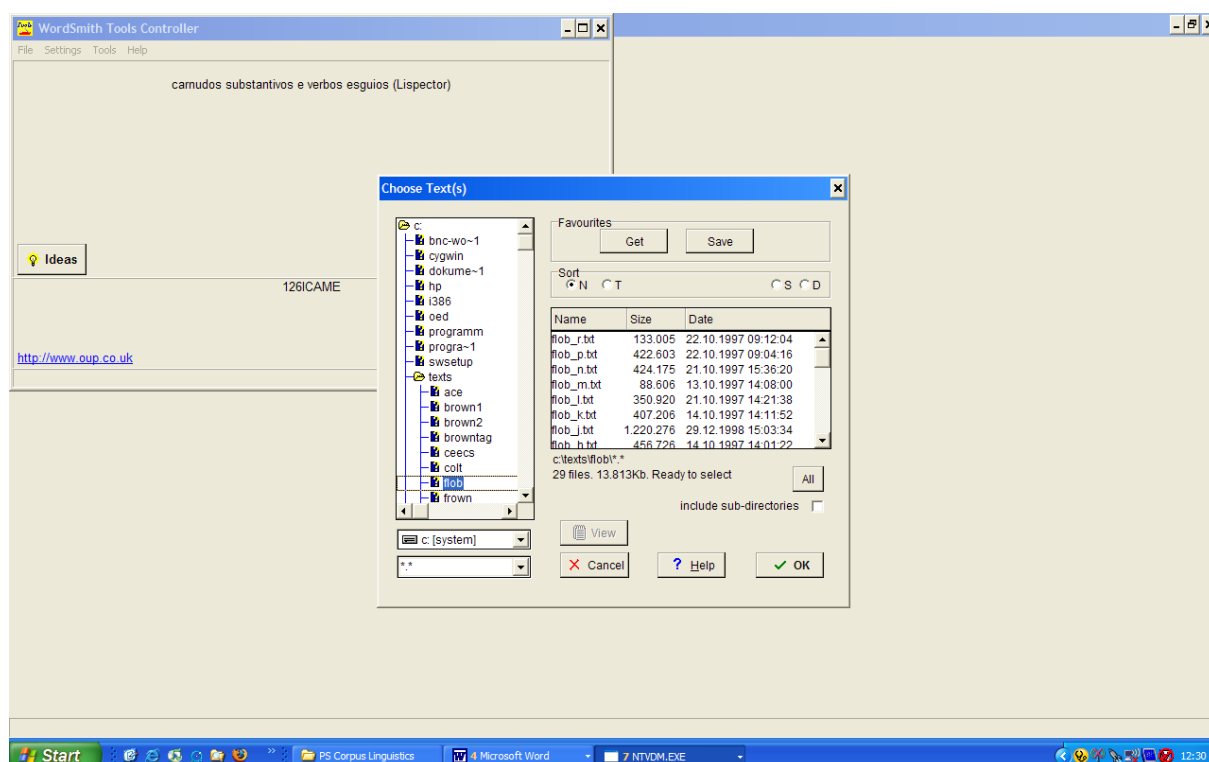
Your next step depends on what you want to do. The two most important options, the WordList function and the Concord function will be introduced here.

(N.B. While almost all corpora can be accessed directly through WordSmith Tools, if you want to work with ICLE (The International Corpus of Learner English), you first have to start the ICLE-program, create your own subcorpus and store it before you can access it through WordSmith; see the ICLE manual)

2) The WordList function

In the left window (“WordSmith Tools Controller”), click on “Tools”. A menu will appear; click on “WordList”. The “WordList” window will appear (which you can now enlarge to full screen size).

Now you have to select the corpus you want to work with. Let us say you want to work with FLOB (a one-million word corpus of written British English from the 1990s). Click on “File” and select “Start” from the menu. A window will pop up, where you select the option “Choose Texts Now”. Another window will pop up, displaying the structure of the files on the computer. Choose the appropriate corpus file. On the corpus computer in the library, the path for FLOB is C:\ICAME\texts\FLOB. When you have clicked your way through to your corpus (FLOB in this case), on the right of the window a list of the text files of which the corpus consists will appear (as the manual will tell you, FLOB – and all related corpora – consists of a number of subcategories, marked by a letter, such as j for science and d for religion, and all the texts belonging to one of these subcategories are stored in one file):



Click on “All” (to select all text categories included in the FLOB corpus, i.e. the whole corpus), and then on “ok”. The “Getting Started” window appears again; this time click on the option “Make a word list now”. Three new windows now pop up, on top of each other: “new wordlist (S)”, “new wordlist (A)”, “new wordlist (F)”:

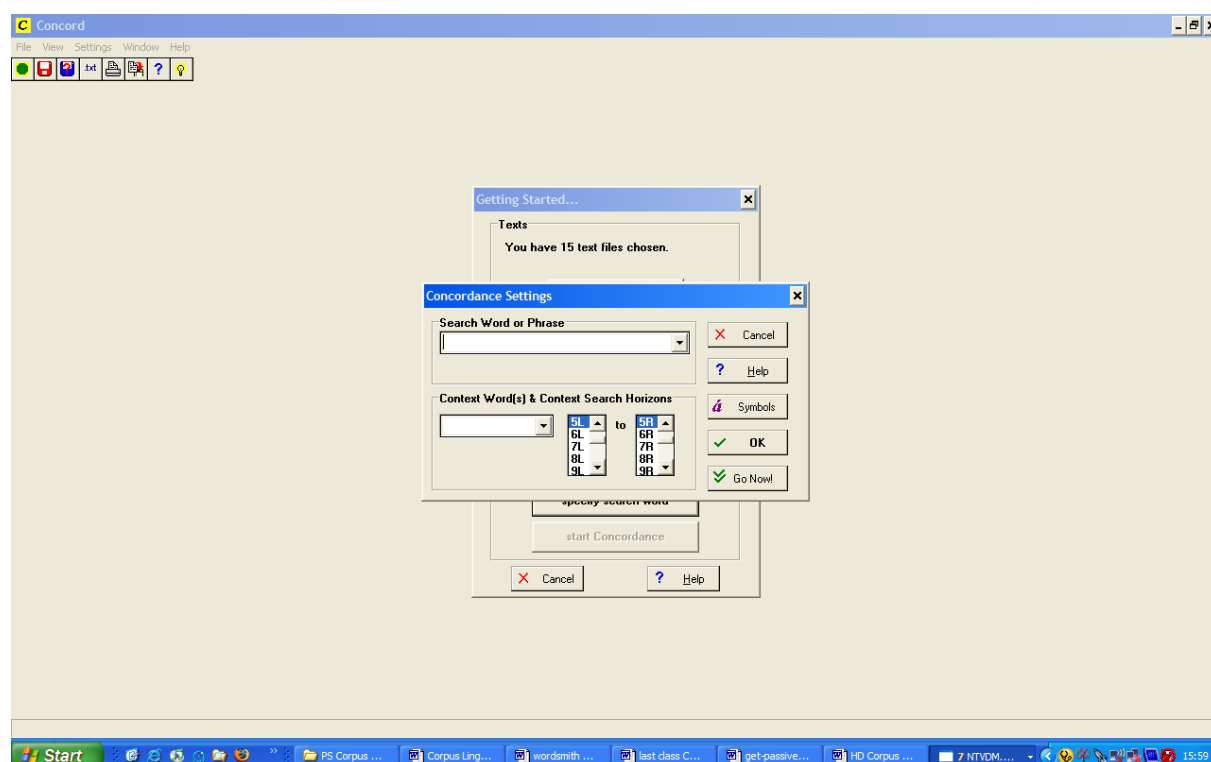
- “new wordlist (S)” gives you frequency information on each of the categories in FLOB (as well as on the whole corpus), most importantly the number of words in each category, but also a great deal of other information, such as the average word and sentence length in each category, the number of words of a certain length etc.
- “new wordlist (A)” gives you an alphabetical list of all the words occurring in the corpus, and the number of occurrences of each single word
- “new wordlist (F)” gives you the same list, only this time sorted as to frequency, with the most frequent words appearing first.

3) The Concord function

The concord function is the most important one for most kinds of linguistic analysis.

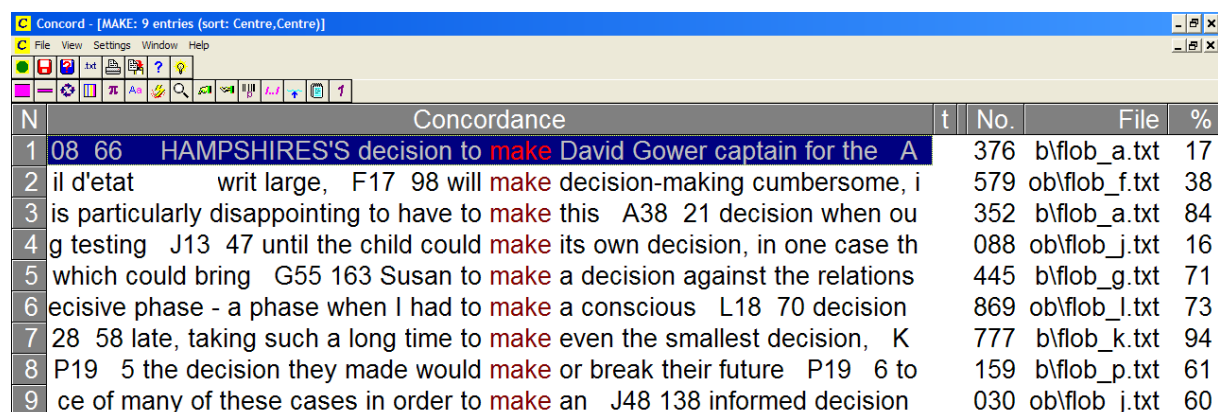
After performing the steps described in “Getting Started” above, in the left window (“WordSmith Tools Controller”), click on “Tools”. A menu will appear; click on “Concord”. The “Concord” window will appear (which you can now enlarge to full screen size).

The selection of the corpus files you want to work with works exactly as described for the WordList function above. When the corpus selection has been completed (by clicking on “All” and then “ok” as described above), the “Getting Started” window appears. Click on “specify search word”, and the following screen (“Concordance Settings”) appears:



You can now enter the word or phrase you would like to investigate under “Search Word or Phrase”. If you then click on “Go Now!”, the concordance lines of your search string appear (as illustrated for *stark* in the Australian Corpus of English in one of the above sections). The searches can be more complex than just for simple words or phrases, however. You can, for example, investigate the occurrence of two words within a certain span (or distance). In that case, you must enter one word under “Search Word or Phrase”, the other under “Content Word(s)”, and in addition indicate the “Search Horizons”, i.e. how many words to the left or right the second word may occur with respect the first. For example, if you are interested in the co-occurrence of the words *make* and *decision* in a span of 5 words, you enter *make* in the

top box, *decision* in the bottom box (or vice versa), and click on “5L” and “5R” (which tells the program to search for *decision* when it occurs either to the left or right of *make* within 5 words). If you run this search by clicking on “Go Now!” you get the following results:



The screenshot shows a software window titled "Concord - [MAKE: 9 entries (sort: Centre, Centre)]". It has a menu bar with "File", "View", "Settings", "Window", and "Help". Below the menu is a toolbar with various icons. The main area displays a concordance table with the following columns: "N", "Concordance", "t", "No.", "File", and "%". The table contains 9 rows of search results, with the word "make" highlighted in red in the concordance text.

N	Concordance	t	No.	File	%
1	08 66 HAMPSHIRE'S decision to make David Gower captain for the A		376	b\fllob_a.txt	17
2	il d'etat writ large, F17 98 will make decision-making cumbersome, i		579	ob\fllob_f.txt	38
3	is particularly disappointing to have to make this A38 21 decision when ou		352	b\fllob_a.txt	84
4	g testing J13 47 until the child could make its own decision, in one case th		088	ob\fllob_j.txt	16
5	which could bring G55 163 Susan to make a decision against the relations		445	b\fllob_g.txt	71
6	ecisive phase - a phase when I had to make a conscious L18 70 decision		869	ob\fllob_l.txt	73
7	28 58 late, taking such a long time to make even the smallest decision, K		777	b\fllob_k.txt	94
8	P19 5 the decision they made would make or break their future P19 6 to		159	b\fllob_p.txt	61
9	ce of many of these cases in order to make an J48 138 informed decision		030	ob\fllob_j.txt	60

If you are interested in the collocation *make a decision*, you are, however, probably also interested in clauses such as *the decision was made* etc., so that you would like to include all inflectional forms of the verb *make* (*make*, *makes*, *made*, *making*) in your search as well as all forms of *decision* (i.e. including *decisions* and perhaps *decision's*). Rather than query the corpus for all combinations of all these forms individually, you can use an asterisk (i.e. the “*” symbol) and ask the program to search for all words beginning with the string *decision* and for all those beginning with the string *ma*. In order to do this, you have to go back to the “Concordance Settings” window by selecting “File” and “Start” (or alternatively, by clicking on the green button in the upper left hand corner), then click on “change search word”, enter *decision** in one of the boxes, and *ma** in the other, and click on “Go Now!”. This search yields 58 hits; below are the first ones (you can ignore the combinations of letters and numbers that sometimes appear in the middle of concordance lines, such as B26 210 in the second line here; they have to do with the format of the corpus, which does not concern us here):

Concordance					t	No.	File	%
1	8	semblance of personal decision-making	to a leader who provides			590	flob\flob_d.txt	86
2	We	B26 210	must remain free to make	our own decisions about ow		809	flob\flob_b.txt	96
3	sting	J13 47	until the child could make	its own decision, in one cas		090	\flob\flob_j.txt	16
4	F14	84	hospital staff in decision-making.	It is increasingly clinician		012	\flob\flob_f.txt	31
5	surement,	encouraging children to make	decisions about when,	H03		832	flob\flob_h.txt	8
6	week. The	Politburo would make	all A05 12	major decisions		188	flob\flob_a.txt	9
7	tion to a	strong A26 38	decision-making	structure, it has kept the p		356	flob\flob_a.txt	57
8	65	affecting a decision in a private	matter"	, and further F27 16		003	\flob\flob_f.txt	61
9	66	HAMPSHIRE'S decision to make	David Gower captain for the			376	flob\flob_a.txt	17
10	at business, and	so, in F02 189	making	decisions, take one anothe		777	\flob\flob_f.txt	4
11	communicated	to central decision-makers,	but the J46 116	costs of		622	\flob\flob_j.txt	57
12	ch could bring	G55 163	Susan to make	a decision against the relati		445	flob\flob_g.txt	71
13	ot be neutral but	A44 12	have to make	political decisions all the tim		439	flob\flob_a.txt	98
14	s'	decisions about public services	may better reflect	J46 22	local r	590	\flob\flob_j.txt	57
15	I	control of resources and decision-making.	F14 91	I see NHS		093	\flob\flob_f.txt	31
16	Council of Ministers	had deferred making	E37 112	a decision on tr		227	flob\flob_e.txt	96
17	ecision	B05 183	IF JOHN Major	hasn't already decided when		791	flob\flob_b.txt	17
18	isions"	(1970: 322).	How, we may	ask, J28 58	did households	446	\flob\flob_j.txt	35
19	37	desirable	is when decisions are	made	by groups that each contain	747	\flob\flob_i.txt	57

Clearly, a number of these hits are irrelevant if you are interested in the collocation *make a decision*. First of all, some words beginning with the string *ma* are not instances of the verb *make* (e.g. *may*, *matter*, *Major*). You would also probably like to exclude the noun (or adjective) *decision-making* and the noun *decision-makers*. In an instance such as number 9 (*HAMPSHIRE'S decision to make David Gower captain*), *make* does not belong to *decision* syntactically; such instances also have to be excluded.

The thinning of instances (i.e. removal of irrelevant instances) works as follows: First, you click on a line showing an irrelevant instance so that it is highlighted and press the delete-button ("entf" on most German keyboards). The line thus treated becomes a lighter shade of grey (If you do that to a line by accident or discover later that you do not want a certain line removed after all, you simply have to press "einf" to get the line back to the original colour). Once you have done this for all irrelevant instances, you click on the "zap"-button (the one that looks like a yellow flash; 7th from the right in the bottom row). The deleted instances then disappear, leaving you with 25 instances in our example:

Concordance				Word No	File	%
N						
8	6	37	desirable is when decisions are made by groups that each contain	J	112.747	\flob_j.txt 57
9	t	A37	154 disregard of the rules and made decisions for himself which wer		92.222	flob_a.txt 83
10	nuine	insights from	his guest. But he made the career decision to	C11	25.969	\flob_c.txt 61
11	ng	D16	5 from the right motive and making the right decision in particular		36.838	flob_d.txt 88
12	parent	or the other	a general 'right' to make	H13	31.802	flob_h.txt 43
13	throoms	for our	home, but we haven't made a	A09	21.296	flob_a.txt 19
14	erson	in her own	right and wanted to make her own	F07	16.838	\flob_f.txt 15
15	iage...	and	P19 5 the decision they made would make or break their futur		44.157	flob_p.txt 61
16	s particularly	disappointing to have to make this	A38	21	93.352	flob_a.txt 84
17	cisive phase - a phase	when I had to make a conscious	L18	70	42.869	flob_l.txt 73
18	cal phenomena.	Thus when a subject makes a decision and his	J51	125	126.139	\flob_j.txt 64
19	28	58	late, taking such a long time to make even the smallest decision,	K	66.778	\flob_k.txt 94
20	day	F03	151 to tell you why you are making the wrong decision and why y		6.576	\flob_f.txt 6
21	P19	5	the decision they made would make or break their future	P19	44.159	flob_p.txt 61
22	ce of many of these	cases in order to make an	J48	138	119.033	\flob_j.txt 60
23	7	128	they finally have someone who makes all their decisions for them,	F	16.358	\flob_f.txt 15
24	learned how to	handle money, how to make	F07	150	16.620	\flob_f.txt 15
25	pulsory viewing	for those involved in making the big	B20	185	49.409	flob_b.txt 73

For easier analysis, you would now like to sort the instances, so that all instances of the same form of *to make* are listed together. Sorting is done as follows: You click the “re-sort” button (4 purple arrows arranged in a circle, the third button from the left in the bottom row), and the following (“Concordance Sort”) window appears:

Concordance				Word No	File	%
N						
8	6	37	desirable is when decisions are made by groups that each contain	J	112.747	\flob_j.txt 57
9	t	A37	154 disregard of the rules and made decisions for himself which wer		92.222	flob_a.txt 83
10	nuine	insights from	his guest. But he made the career decision to	C11	25.969	\flob_c.txt 61
11	ng	D16	5 from the right motive and making the right decision in particular		36.838	flob_d.txt 88
12	parent	or the other	a general 'right' to make	H13	31.802	flob_h.txt 43
13	throoms	for our	home, but we haven't made a	A09	21.296	flob_a.txt 19
14	erson	in her own	right and wanted to make her own	F07	16.838	\flob_f.txt 15
15	iage...	and	P19 5 the decision they made would make or break their futur		44.157	flob_p.txt 61
16	s particularly	disappointing to have to make this	A38	21	93.352	flob_a.txt 84
17	cisive phase - a phase	when I had to make a conscious	L18	70	42.869	flob_l.txt 73
18	cal phenomena.	Thus when a subject makes a decision and his	J51	125	126.139	\flob_j.txt 64
19	28	58	late, taking such a long time to make even the smallest decision,	K	66.778	\flob_k.txt 94
20	day	F03	151 to tell you why you are making the wrong decision and why y		6.576	\flob_f.txt 6
21	P19	5	the decision they made would make or break their future	P19	44.159	flob_p.txt 61
22	ce of many of these	cases in order to make an	J48	138	119.033	\flob_j.txt 60
23	7	128	they finally have someone who makes all their decisions for them	F	16.358	\flob_f.txt 15
24	learned how to	handle money, how to make	F07	150	16.620	\flob_f.txt 15
25	pulsory viewing	for those involved in making the big	B20	185	49.409	flob_b.txt 73

Concordance Sort

Main Sort ...

5L
4L
3L
2L
1L

then by ...

Centre
1R
2R
3R
4R

finally by ...

Set
File
Len
Tag
Distan

Sort

all

Case Sensitive

Ascending

A with A, C with C etc.

When you click on “Centre” in the category “Main Sort”, your results will be sorted according to the form of *make* (i.e. the central word). You can also sort the results according to the words occurring to the left or right of the search term, sort them on several levels (for

example first according to the search term and then according to the word to the right of it), and sort them according to a number of other criteria, most importantly the file names.

N	Concordance	Word No	File
8	measurement, encouraging children to make decisions about when, H03 9	5.832	flob_h.txt
9	erson in her own right and wanted to make her own F07 166 decisions. A	16.838	\flob_f.txt 1
10	learned how to handle money, how to make F07 150 decisions for herself	16.620	\flob_f.txt 1
11	e. We B26 210 must remain free to make our own decisions about own d	64.809	flob_b.txt 9
12	nnnot be neutral but A44 12 have to make political decisions all the time i	108.439	flob_a.txt 9
13	28 58 late, taking such a long time to make even the smallest decision, K	66.778	\flob_k.txt 9
14	ce of many of these cases in order to make an J48 138 informed decision	119.033	\flob_j.txt 6
15	cisive phase - a phase when I had to make a conscious L18 70 decision	42.869	\flob_l.txt 7
16	parent or the other a general 'right' to make H13 149 decisions about a pa	31.802	flob_h.txt 4
17	s particularly disappointing to have to make this A38 21 decision when ou	93.352	flob_a.txt 8
18	P19 5 the decision they made would make or break their future P19 6 to	44.159	flob_p.txt 6
19	e a week. The Politburo would make all A05 12 major decisions on	10.188	flob_a.txt
20	cal phenomena. Thus when a subject makes a decision and his J51 125	126.139	\flob_j.txt 6
21	7 128 they finally have someone who makes all their decisions for them, F	16.358	\flob_f.txt 1
22	and at business, and so, in F02 189 making decisions, take one another i	4.777	\flob_f.txt
23	ng D16 5 from the right motive and making the right decision in particular	36.838	flob_d.txt 8
24	day F03 151 to tell you why you are making the wrong decision and why y	6.576	\flob_f.txt
25	pulsory viewing for those involved in making the big B20 185 decision.	49.409	flob_b.txt 7

This screen shows the results sorted first according to the central word (“Center” in “Main Sort”) and secondly according to the words on the left of *make* (“1L” selected under “then by”, then “ok”).

If you want to look at more context, you click on the line in question and then click on the pink “grow” button, at the very right in the bottom row. (To get back to the original concordance lines, you can later click on the pink “shrink” button right next to it.)

In order to save the results for later analysis, click on “File”, and then “Save As”, select a drive and a folder on your computer and click on “ok”.

A (sizeable) manual explaining all the functions of WordSmith in detail is available in the library next to the corpus computers (It can also be downloaded from <http://khnt.hit.uib.no/icame/manuals/wsmanual.pdf>). A help function is also directly available in the program, and can be accessed from the window “Word Smith Tools Help” (which pops up in front of all others when WordSmith is started).

3) Exercises (I and II)

For step-by-step solutions click here (password required).

To obtain the password, please write an email to Nadja.Nesselhauf@urz.uni-heidelberg.de.

I Using the WordList function of WordSmith

- 1) What are the 5 most frequent words in Frown?
- 2) How frequent is the word *ingredient* in Frown? Is the singular or the plural form more frequent?
- 3) The file FROWN_P.txt contains excerpts from the category “Romance and love stories”. How many words (“tokens”) does this category consist of in the corpus?
- 4) Compare the 5 most frequent words in Frown to the 5 most frequent words in FLOB.
N.B. When you select new texts / a new corpus to work with in WordSmith (“File” – “Start”, then “Change Selection” in the “Getting Started” window), you first have to clear your previous selection by clicking on “clear previous” in the “Choose text(s)” window”! Otherwise, the new selection of texts is added to the old selection.
- 5) If x is the year in which you were born, what is the xth frequent word in FLOB and how frequent is it? (For example, if you are born in 1982, which is the word listed under number 1982 in “new wordlist (F)”)? Please also state what x is.

II Using the Concord function of WordSmith

- 1) Does the word *gamut* occur in Frown? If so, what are the words occurring to the right of *gamut*?
- 2) Select a word that starts with the same letter as your first name and that you think is quite frequent in English.
 - a) How frequently does it occur in Frown?
 - b) What is the word occurring most frequently to the left of the word (use the sorting function) and how often does it occur?
- 3) The word *conjugal* appears once in Frown. Give the full sentence in which it occurs (The sentence starts with the word *In*; ignore the 6-digit combinations of letters and numbers interspersed in the text).
- 4) What is the adjective occurring most frequently directly to the left of the word *factor* (use the sorting function) and how frequently does it occur?
- 5a) What is the easiest way of finding out whether and how often the collocation *take a trip* (including instances such as *When we were taking that trip...*, *The trips we are going to take...* etc.) occurs in Frown (select a distance of 6 words to both sides)?
 - b) How many results does your search / do your searches yield?

c) Are all of them instances of the collocation *take a trip*? If not, give one example of a result that is not an instance of this collocation. Remove this instance (i.e. the whole line) from the concordance.

4) How to conduct linguistic analyses on the basis of corpora: two examples

Example 1: Australian English vocabulary

Let us imagine you are interested in whether Australian English vocabulary tends to be more similar to British or to American English.

First, you formulate some more precise questions or hypotheses that you would like to investigate:

- Does Australian English exclusively use vocabulary items from either British or American English or does it use items from both varieties?
- If it uses items from both varieties, does it usually make a choice for a given pair of items or does it in addition mix items denoting the same concept from both varieties?
- If items from both varieties are used, can a dominance of either British English or American English vocabulary be observed?
- If Australian English uses terms from both varieties denoting the same concept in British and American English, can a difference in usage of these two items be observed?

In a next step you consult a reference book, for example Crystal's *Encyclopedia of the English Language*, for differences of vocabulary in British and American English (1995: 309) and randomly choose 5 pairs of items from among those which are said to be used exclusively in one of the two varieties: *flashlight* (AmE) – *torch* (BrE), *freeway* (AmE) – *motorway* (BrE), *sidewalk* (AmE) – *pavement* (BrE), *elevator* (AmE) – *lift* (BrE), *diaper* (AmE) – *nappy* (BrE).

Then you choose your corpus. For Australian English, the Australian Corpus of English (ACE) is available to you. This corpus contains written English of different text types (you need to keep in mind, therefore, that your study has two major caveats: first, your investigation is limited to only a very small part of the Australian English vocabulary, and secondly, your investigation is based on written language only). The software you choose for the investigation is WordSmith; you start by searching for *flashlight** and *torch** (the asterisk is necessary to include the plural forms), and get the following results:

Concordance					Se	Word No.	File	%
N	1	the pitch darkness	they were running their flashlights	back and forth over the tail asse		16.653	ace_b.txt	30
	2	t stairs, and had climbed up	to the tail with flashlights.	In the pitch darkness they were		16.644	ace_b.txt	30

Concordance					t	T	Word No	File	%
N	1	ibious	landing operation entitled "Beacon Torch"	will be launched from the Tripoli.			5.252	e\ace_n.txt	43
	2	e	was an operation code named "Beacon Torch".	Bong! Bong! Bong! "Hands to a			5.018	e\ace_n.txt	41
	3	sound	of footsteps and a passing flash of torchlight	assured me that the local night			32.915	e\ace_g.txt	28
	4	f	Larry's crust. She set it to wide beam to torch	the hut in which she and Katrina ha			4.331	\ace_m.txt	36
	5	ty	pewter thimbles for \$7, moulded plastic torches	just like the one held by the stat			19.571	e\ace_a.txt	22
	6	an	who came towards me carrying a huge torch.	K23			14.138	e\ace_k.txt	58

The next step is to investigate whether the occurrences are really occurrences of the concept in question. In the case of *flashlights*, the co-occurrence of the words with *pitch darkness* indicate that some kind of lamp is indeed referred to. In the case of *torch*, the first two occurrences appear to be part of a proper name (that refers to some kind of operation) and therefore need to be disregarded. The third instance that has been thrown up by the search is neither an instance of *torch* nor of *torches*; the fourth is an instance of the verb *to torch*. These two instances need to be disregarded. The final two instances actually are instances that seem to refer to a portable lamp. When you take a closer look at the files in which the instances

occur (under “File” on the right of the window), it turns out that *torch* occurs in two different categories, a and k (and therefore in two different texts), whereas the two instances of *flashlight* occur in the same category (category b). If you look at the context (you might want to “grow” the line to receive more context), it turns out that they occur in the same text, in two consecutive sentences. A common notation for this is 2 (1), i.e. two occurrences in one text. You then perform searches in the same fashion for the other words: *freeway**, *motorway**, *pavement**, *sidewalk**, *elevator**, *lift**, *diaper**, *napp** (N.B. a search for *nappy** would not throw up the plural form in this case, so you either have to perform two searches, for *nappy* and *nappies*, or a search for *napp**) These searches give you the following output:

N	Concordance	t	Word No	File %
1	m for the next 20km. 3. When exiting the freeway, stay in the right-hand lane until 5	26.769	ace_b.txt	48
2	Daily Mirror - 4 August 1986 NEW FREEWAY HITS THE PITS! Drivers	8.304	ace_a.txt	9
3	u successfully negotiate your way onto the freeway and pull out into the right lane to	44.253	ace_b.txt	79
4	Tow trucks were pulled off the side of the freeway. Their cabin lights went round an	17.380	ace_p.txt	78
5	such serious problems with road capacity, freeways, pollution, congestion, excessive	67.280	ace_g.txt	58
6	hought the recently-completed F4 Western Freeway would save them hours of travel	8.327	ace_a.txt	10
7	k learner. This is what I have picked up. - Freeway driving: 1. Stay in the right-hand l	26.722	ace_b.txt	48

A close inspection of the contexts of these instances leads you to the conclusion that they are all instances of the relevant concept.

There are no results for *motorway**.

N	Concordance	t	T	Word No	File %
1	eways, empty cartons kicked along the pavements. Dog turds. Not*no so far, p	1.653		ace_k.txt	7
2	y, should be enforced for riding on the pavement. Suggest+ions ranged from h	5.220		ace_f.txt	9
3	who have found themselves out on the pavement much richer for having poor	44.539		ace_b.txt	80
4	ne. She lifted her hand as I reached the pavement, and closed the door. The n	670		ace_k.txt	3
5	they stayed inside now. Heard from the pavements. Her street was close yet re	1.725		ace_k.txt	7
6	and roll-neck sweaters clomping on the pavement. Standing at the kerb blowing	18.276		ace_k.txt	75
7	many people of all ages congesting the pavement in front of Myer store. Excite	2.261		ace_k.txt	9
8	ew a dirty felt of spent leaves across the pavement. Where there was a kind of s	1.759		ace_k.txt	7
9	sed-up faces, staring down at the grey pavement, avoiding each other as they	18.330		ace_k.txt	75
10	rds of suburban boys drifted along the pavements chomping hamburgers and f	101.815		ace_g.txt	89
11	ntil there was no verge, only the strip of pavement. Traffic crowded towards ligh	202		ace_k.txt	1

N	Concordance	t	T	Word No	File %
1	r of artists painting at easles among the sidewalk throng. His stay in Europe h	16.510		ceace_k.txt	67
2	and snow and the mountain lying fir; at sidewalk stalls he ate herring in rolls, s	17.453		ceace_k.txt	71
3	nd people. There are open air markets, sidewalk cafes and bazaars in every noi	18.515		ceace_e.txt	33

All instances of *sidewalk** and *pavement** appear to refer to the relevant concept (again, if you are unsure about certain instances you can retrieve more context to decide).

There are no results for *elevator*.

For *lift**, you get 66 results. Glancing through them, you notice that many of these are instances of the verb *lift* rather than of the noun. One easy way to get rid of a number of the verb-instances is to perform the search again, only this time by excluding the form *lifted* (in the window “Concordance Settings”, type in *lifted* in the “but excluding”-box, which appears whenever you type in a string containing an asterisk in the “Search Word or Phrase” box). This leaves you with 40 instances. The form *lifting* is not relevant to your investigation either,

so you sort the instances according to the central word, and remove all instances of *lifting*. As you glance through the remaining instances, you notice that many instances of the verb *lift* are preceded by *to*. You perform another sort, this time according to the word to the left of the search term, and remove the instances of *lift* preceded by *to*, which leaves you with the following 24 instances:

Concordance			
N	t	Word No.	File %
1	up. Two crisps blue-garbed apes stepped out of a lift . Mental health and social adjust+ment radiated	9.326	ace_m.txt 77
2	utes shaking his head in amazement. What next - a lift to the top of Ayers Rock? Then it struck him tha	2.198	\ace_f.txt 4
3	orse-trainer happened to be going my way, I had a lift home on a trotting spider. One morning, having	22.026	\ace_g.txt 19
4	es provided a sure source of income - it installed a lift from the dun+geons to the top floor, which also	35.325	\ace_g.txt 30
5	m the north coast to Sydney Johnston had given a lift to his mate's wife. It had taken two days. The si	17.667	\ace_p.txt 80
6	also boasts the cheapest daily entry fee - \$3 - and lift tickets - around \$12.60 adult, \$6.40 child. E19	16.293	\ace_e.txt 29
7	ll volcanic peak with a sliver of beach. As the cloud lifts he can see out to the reef. It's always a reass	18.935	\ace_k.txt 78
8	nd Soviet Union divert vast resources which could lift the burden of world poverty, to the wasteful, se	18.320	\ace_d.txt 53
9	abortionist's, her attempted suicide, her failed face-lift , Dorothy Rainbow's massacre of her children a	91.267	\ace_j.txt 82
10	t their duties the following morning in readiness for lift-off , while the indigenous inhabitants of Le Gard	5.281	ace_m.txt 44
11	raffic congestion, parking hassles and queueing for lift tickets and chairlifts. It's little wonder then that	14.656	\ace_e.txt 26
12	od involved in human history (10-19)?MEDITATE I lift my eyes to the hills From whence does my help	30.293	\ace_d.txt 87
13	ouple of thousand people start cheering my name it lifts my performance; it spurs me on. And I always l	22.229	\ace_a.txt 25
14	ow motel for five days with all meals (exceptlunch), lift tickets, lessons, ski hire, and bus travel to the s	16.219	\ace_e.txt 29
15	in+creased faster than prices have risen, you must lift productivity to survive. G76 The Bulletin - July 1	109.900	\ace_g.txt 96
16	ut the role of religion in Australian society. It should lift the debate by increasing the information on the	12.514	\ace_d.txt 36
17	irrigation will never be realised, but it is a sight that lifts your heart, this great water in a dry land. Ther	61.902	\ace_a.txt 68
18	gently out into the corridor and was wafted up the lift well, which acted as a funnel to disperse the fra	35.447	\ace_g.txt 30
19	egan to push the couch and her numb body into the lift . "Candidly, Ted," she heard Frank say, as the d	9.933	ace_m.txt 82
20	m the cocoon. A melodious tone sounded from the lift . "Ah, there we are now. There's no need for an	9.836	ace_m.txt 81
21	ape of great beauty. These skiers are far from the lift queues of the resorts, for they are cross-countr	14.350	\ace_e.txt 25
22	with, ease US trade problems, improve world trade, lift commodity prices and ease Third World debt pr	70.583	\ace_a.txt 78
23	yone report for duty at 0400 ships time and we will lift off at 0500, which will be about mid-morning pla	5.153	ace_m.txt 43
24	cal expansion by West Germany and Japan would lift world economic growth, ease US trade problem	70.572	\ace_a.txt 78

You go through these one by one and decide whether they are instances of the relevant concept. While doing that, you notice that it is not only the verb *to lift* that has to be excluded from the count, but also other meanings of the noun *lift*, in particular the meaning as in *to give sb. a lift* and those instances referring to a *ski lift* (if you are uncertain whether there is a British-American difference for this term, consult a dictionary). To decide whether certain instances refer to a ski lift or a lift in a building, you will need to look at more context in some cases. In the end, you are left with 6 instances (numbers 1, 2, 4, 18, 19, 20).

There are no results for *diaper**.

Concordance			
N	t	Word No.	File %
1	m the hospital. Elinor had collected the big nappy bag Susan had carried with her. It als	2.615	celace_s.txt 13
2	d be pressed into use decorating disposable- nappy ads long before they could be attach	105.593	celace_g.txt 93
3	ed Darren had cockroach infestations in his nappy and his stomach had been scalded b	56.255	celace_a.txt 62
4	world where the cherubs cried and wet their nappies , where bunches of grapes moved a	50.312	celace_g.txt 43
5	allegedly suffered scalding from wearing wet nappies for too long. Penrith Court was tol	56.297	celace_a.txt 62

All 5 instances of *nappy* appear to be instances of the concept in question.

To get an overview over your results, you enter them in a table.

words:	occurrences of British word:	occurrences of American word:
torch : flashlight	2	2 (1)
motorway : freeway	0	7

<i>pavement</i> : <i>sidewalk</i>	11	3
<i>lift</i> : <i>elevator</i>	6	0
<i>nappy</i> : <i>diaper</i>	5	0
total:	24	12

As with the first example (*flashlight* and *torch*), you could check whether in those cases with several occurrences, the words occur in the same text or not. As all the texts of one category are stored in one file in this corpus, this is difficult to find out in some cases, however. If the instances occur in different categories, they also appear in different texts. If they appear in the same category, a good indication is the word number (which is given next to the file name in the concordance line; a word number such as 5,212, for example, means that in the category in question the search term is the 5,212th word). As the texts in the Australian Corpus of English are only around 2,000 words long (see the corpus manual), two instances that have word numbers more than 2,000 words apart are unlikely to occur in the same text. If they have word numbers that are apart less than 2,000 words, one way to find out is to open the file with a word processor and use the search function of that program. For the purposes of this investigation, you judge this as unnecessary, however.

What do these results tell you with respect to your research questions then?

First, Australian English does not exclusively use either British or American vocabulary. Secondly, the results indicate that in some cases, in Australian English either the British or the American word is used exclusively (but the results are no conclusive proof: a larger corpus might throw up instances of the other variety): In the case of *freeway* – *motorway*, only the American word occurs in the corpus, in the case of *nappy* – *diaper* and *lift* – *elevator*, only the British word occurs. This might lead to the hypothesis (which in turn might be the basis for further research) that words relating to traffic and cars might be dominantly American in Australian English. In some cases, however, Australian English uses both the American and the British term. In the case of *torch* and *flashlight*, the words might be used with similar frequencies (though the numbers are too small for any firm conclusions). In the case of *pavement* and *sidewalk*, *pavement* appears to be the dominant variant. Moreover, it appears from the results that *sidewalk* is only used in modifying function (*sidewalk throng*, *sidewalk stalls*, *sidewalk cafes and bazaars*). None of the instances of *pavement* occurs in this function. So this might be a case where the words from the two varieties have undergone a functional distinction in Australian English.

Although the words investigated are of course too few for any definite conclusions, the results (see the totals in the table) indicate that while British vocabulary is dominant in Australian English, there has also been some influence of American English. (In a larger study, you would of course not only include more items, but also contrast your results with those recorded in relevant literature on Australian English).

Example 2: Present perfect and simple past in British and American English

You are interested in differences of the use of the present perfect and the simple past in British and American English. In Swan, *Practical Guide to English Usage*, you find that in American English the simple past can be used with adverbs such as *just*, *already*, *yet*, *ever*, and *before*, whereas in British English only the present perfect is possible (1995: 423). First, you formulate some questions or hypotheses on what precisely you want to find out:

- Does the simple past not occur at all with these adverbs in British English?
- Does American English allow only the simple past, or also the present perfect?
- If American English allows both the simple past and the present perfect, which is more frequent (and how much more frequent exactly)?
- If American English allows both, when is what tense / aspect used (i.e. does the choice depend, for example, on the verb of the clause in question, are there slight differences of meaning depending on which tense / aspect is used etc.)?

Then you choose corpora on which to base your analyses. Since you are interested in Present Day English, and in American and British English, two comparable corpora of the two varieties in question would be ideal. From the list provided on this website you glean that the two most suitable corpora for your investigation are FLOB and Frown. FLOB and Frown are two matching corpora of British and American English of the 1990s, which contain different written text types. (This is one important caveat of your study, then, which should be kept in mind: you are only investigating written English and not spoken English).

The software you chose for your investigation is WordSmith. You decide to start your investigation with the adverb *just*. In order to answer the first three questions, you first look at the relative occurrence of the simple present and the present perfect with *just* in both varieties. First you search for *just* in FLOB and Frown: In FLOB you find 1107 occurrences, in Frown 1124. Too many to look at individually. What can you do?

One option is to choose a few verbs which are frequent and have the program look up occurrences of these verbs which occur next to *just*. You choose the verbs *be*, *do*, *go*, *give*, *hear*, and *say*. The next point you have to consider is where in relation to *just* and in which form the verb usually occurs in a clause when *just* refers to the past. *Lucy has just called* is the example given in Swan (1995: 419) for British English; *Lucy just called* the example for American English. So the form of the verb occurs to the right of *just* (**Lucy called just* or **Lucy just has called* or **Just Lucy has called* are very unlikely to occur with the 'very recently' meaning of *just*). As the present perfect is formed with the past participle and the simple past with the past tense form (which are not always identical), and as the past tense has two forms for the verb *to be*, the following forms will have to be investigated: *been*, *was*, *were*; *done*, *did*; *gone*, *went*; *given*, *gave*; *heard*; *said*. You first investigate the usage in FLOB (the British corpus).

Just been yields 10 instances:

N	Concordance	t	T	Word No.	File %
1	remaining primeval forests had just been E22 155 murdered.	53.768	ts\fløb\fløb_e.txt	57	
2	wledge 1, in spite of what has just been said, is a form of J33	80.403	xts\fløb\fløb_j.txt	41	
3	nion Mr Alan Cooper, who has just been cleared of B13 186 i	31.868	ts\fløb\fløb_b.txt	47	
4	dramatic, somehow, for having just been K14 123 abandoned.	33.395	ts\fløb\fløb_k.txt	47	
5	7 57 that." A17 58 Perhaps he just been sic! unlucky. It was	40.963	ts\fløb\fløb_a.txt	37	
6	. "You look hot." P24 159 "I've just been working out," she ex	58.481	ts\fløb\fløb_p.txt	82	
7	oks," he began. P13 102 "I've just been out to measure up s	30.796	ts\fløb\fløb_p.txt	43	
8	it of fishing. My youngest lad's just been called up, so I K16 1	38.351	ts\fløb\fløb_k.txt	54	
9	is E23 49 time last year she'd just been offered a place at co	54.980	ts\fløb\fløb_e.txt	58	
10	L03 97 And since then there's just been the two of us. You're	5.917	xts\fløb\fløb_l.txt	10	

Just was and *just were* yield 0 instances.

Just gone yields 2 instances:

N	Concordance	t	No.	File %
1	dealers in club memberships has just gone bust (can you F21 13	050	texts\fløb\fløb_f.txt	47
2	casionally, I G17 88 must have just gone into the Troc for a drink	213	exts\fløb\fløb_g.txt	21

Just went yields 0 instances.

Just done yields 1 instance:

N	Concordance	t	No.	File %
1	our, because they'd L03 185 just done the news headlines	876	texts\fløb\fløb_l.txt	11

Just did yields 0 instances.

Just given yields 1 instance:

N	Concordance	t	T	No.	File %
1	9 30 through his hair. "But Dad's just given me a message saying P29	.195	fløb\fløb_p.txt	97	

Just gave yields 0 instances.

Just heard yields 0 instances.

Just said yields 4 instances:

N	Concordance	t	T	No.	File %
1	Kramer, then realized what he had just said. L05 116 And he wa	.999	ts\fløb\fløb_l.txt	19	
2	, unable to remember what she had just said, let B21 43 alone what h	.291	s\fløb\fløb_b.txt	75	
3	211 champagne bottle, you know. I just said that to tease Oster. I fell	.317	s\fløb\fløb_n.txt	41	
4	M06 201 "I thought you just said she was M06 202 beautif	.654	s\fløb\fløb_m.txt	98	

The next step is a careful consideration of the occurrences. First of all, the instances of the past perfect (*had just* + past participle) have to be excluded, such as *had* / *'d just been*. In one case, *Perhaps he just been unlucky*, it cannot be decided whether this is an instance of past perfect or present perfect, so it has to be disregarded. One instance, *having just been abandoned* is a participle clause, which has to be disregarded as well (as participle clauses are

non-finite and therefore tenseless). Then, the semantics of *just* have to be considered. In two cases at least (*there's just been the two of us* and *I just said that to tease Oster*), *just* clearly means 'only' and not 'very recently', so these instances have to be disregarded as well. For the instances where the concordance lines do not provide enough context to exclude the 'only'-meaning of *just* with some definiteness, you have to look at more context to decide. The instance *has just gone bust*, for example, could be both an instance of the 'only'-meaning and of the 'very recently'-meaning of *just*, but the context seems to support a 'very-recently'-reading. In *I must have just gone*, however, *just* turns out to be an instance of the 'only'-meaning.

The manual sifting thus leaves you with the following results for FLOB:

forms:	number of occ.
<i>just been</i>	5
<i>just was / were</i>	0
<i>just done</i>	0
<i>just did</i>	0
<i>just gone</i>	2
<i>just went</i>	0
<i>just given</i>	1
<i>just gave</i>	0
<i>just heard</i> (present perfect)	0
<i>just heard</i> (simple past)	0
<i>just said</i> (present perfect)	0
<i>just said</i> (simple past)	1

In FLOB, therefore, of the 9 instances of the 6 verbs with the adverb *just*, 8 take the present perfect, and one the simple past. When we look at the instance in the simple past (*I thought you just said that she was beautiful*) more closely, however, it turns out that it is most likely a case of a backshifted present. As our numbers are small, we cannot, therefore, say with confidence whether the simple past tense does not occur with *just* in British English (except as backshifted present) or whether it is just much less frequent than the present perfect. What we can say is that our results indicate that the present perfect is clearly predominant in (written) British English.

You then perform the same searches in Frown, with the following output:

Just been:

Concordance				Se	T	d No.	File	%
1	tional beauty, one who had G36 112 just been on safari in Africa "with 14					87.516	\frown_g.txt	48
2	t Professor Byron M04 70 Snipes had just been published in the avant-gar					8.268	\frown_m.txt	57
3	28 Warner Brother's cartoon that had just been smashed by a frying N28 2					67.148	\frown_n.txt	93
4	rticularly in light K15 203 of what I've just been remembering about my mo					36.568	\frown_k.txt	52

Of these, three are instances of the past perfect; the last one is an instance of the present perfect.

The searches for *just done* and *just did* yield no instances.

Just gone is represented by one instance, which again is an instance of the past perfect, however:

N	Concordance	t	T	No.	File %
1	e says. "Here's this woman who had just gone all out for E01 105 the whole ra			1.242	frown_e.txt 1

There is one instance of *just went*:

N	Concordance	t	T	No.	File %
1	didn't say anything at first either, K27 87 just went on shoveling the food in, and I wa			.724	n\frown_k.txt 9

This, however, is an instance of the 'only'- meaning of *just* and therefore has to be disregarded.

Just given has one instance, which is, however, an instance of the past perfect:

N	Concordance	t	T	No.	File %
1	e outcome of the orders N28 108 he had just given, there was no way to tell. He w			.000	n\frown_n.txt 94

Just gave:

N	Concordance	t	T	No.	File %
1	y mother's letters came more often and just gave me news of G48 154 our do			.428	wn\frown_g.txt 64
2	in and took the poor thing and the guy just gave it up. The L22 117 asshole c			.232	wn\frown_l.txt 90
3	E16 113 Harmon say, "But Hap, I just gave him my E16 114 permission.			.598	wn\frown_e.txt 42
4	5 wife talk, mumbo-jumbo talk. Maybe I just gave you a great big P28 126 earf			.012	wn\frown_p.txt 96

Here, the instances of *just* meaning 'only' have to be excluded. If you look at more context, it becomes clear that only instance number 3 is an instance of the 'very recently' meaning.

There are two instances of *just heard*, both instances of the simple past:

N	Concordance	t	T	No.	File %
1	ok Buck's hand. N03 158 "I just heard what happened. You make			6.722	\frown_n.txt 10
2	Thud. A16 83 That sound you just heard was the Highland High foot			.443	\frown_a.txt 35

Just said yields two instances:

N	Concordance	t	T	No.	File %
1	118 so when she realized what he had just said. N14 119 "Owl, di			.822	wn\frown_n.txt 47
2	ce: "For those of you at home, he just said C03 98 clucking - 'We			6.095	wn\frown_c.txt 14

While the first of these is an instance of the past perfect, the second actually seems to be one of the 'very recently' meaning if the context is inspected (*he just said clucking* – '*We clucking appreciate it*').

forms:	number of occ.:
<i>just been</i>	1
<i>just was / were</i>	0
<i>just done</i>	0
<i>just did</i>	0

<i>just gone</i>	0
<i>just went</i>	0
<i>just given</i>	0
<i>just gave</i>	1
<i>just heard</i> (present perfect)	0
<i>just heard</i> (simple past)	2
<i>just said</i> (present perfect)	0
<i>just said</i> (simple past)	1

In Frown, of the 5 occurrences of the 6 verbs 4 therefore are in the simple past and one in the present perfect. As the numbers are small, we have to be careful with our conclusions. It seems from the results, however, that in American English the simple past is indeed predominantly used with *just* but that the present perfect also occurs. Due to the small numbers, the fourth question asked at the beginning cannot be answered on the basis of this analysis only.

What you have to take into account when performing a corpuslinguistic analysis

As the above analyses have shown, there are certain points that you need to take into account in each corpus-based analysis:

- It is essential that you know what exactly is in the corpus (e.g. what kinds of text, what variety, how many words a text contains, what the different files contain, what the abbreviations in the file names stand for etc.); so always have a look at the corpus manual before you start your work
- If you use different corpora for an analysis, find out to what degree they are comparable (corpus size, variety, time span covered, text types) etc.
- You need to make sure that your search (or searches) actually finds all occurrences of the language item in question (for example all inflectional variants)
- After performing a search with corpus software, you have to check whether all results are actually relevant for the investigation in question (and if not which ones have to be excluded)
- You need to check how the item in question is distributed in the corpus (e.g. whether it only occurs in certain periods / text types etc.); it is also relevant from how many different texts (or text categories) the solutions come from (i.e. if an item is very frequent, but occurs only in one or two texts, this does not mean that it is frequent in the variety / text type etc. investigated)
- If an item does not occur in a given corpus, this does not necessarily mean that it does not occur in the variety / text type / period etc. represented in the corpus, especially if the corpus is fairly small
- Small differences in frequency can be a matter of chance and therefore need to be interpreted prudently
- In the interpretation of your results, be aware of what you have actually investigated (e.g. only written or only spoken language, only certain instances of the phenomenon in question etc.)

5) Exercises (III)

Exercise III.1:

Continue the investigation started in Example 2. Choose another adverb from the list given in Swan (see above, first paragraph in Example 2) and see whether the results for this adverb can contribute further to answering the four research questions. Give not only your results but state how you have arrived at them.

Exercise III.2:

For step-by-step solutions [click here](#) (password required).

To obtain the password, please write an email to Nadja.Nesselhauf@urz.uni-heidelberg.de.

You are interested in the recent development of *whom* in British English. On the basis of the corpora LOB (1961, British English) and FLOB (1991, British English), you attempt to answer the following two questions:

- a) Was there an increase or a decrease (or neither) of *whom* in the course of these 30 years?
- b) Does the (possible) development affect those instances of *whom* after a preposition (*to whom*, *with whom* etc.) to the same degree as those without?

(N.B. LOB and FLOB contain the same text types and the same number of words, so that the results are directly comparable)

(N.B. when analysing the concordance lines, ignore the letter-number complexes that sometimes interrupt the text; these may also affect the sorting function)

Describe briefly what you have done in order to find answers to these questions and present your results in the following table as well as verbally:

	<i>whom</i> (total)	<i>whom</i> (after preposition)	<i>whom</i> (without preposition)
LOB (1961)			
FLOB (1991)			

6) Where to find further information on corpus linguistics

Other web-based introductions to corpus linguistics:

<http://www.georgetown.edu/faculty/ballc/corpora/tutorial.html>

(Concordances and Corpora Tutorial by Catherine Ball at Georgetown University.)

<http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm>

(By McEnery & Wilson, intended as a supplement to their 2001 book.)

General books on corpus linguistics:

Biber, Douglas; Conrad, Susan; Reppen, Randi (1998). *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: CUP.

Kennedy, Graeme (1998). *An Introduction to Corpus Linguistics*. London & New York: Longman.

McEnery, Tony; Wilson, Andrew (2001). *Corpus Linguistics*. Edinburgh: EUP.

Meyer, Charles F. (2002). *English Corpus Linguistics. An Introduction*. Cambridge: CUP.

Partington, Alan (1998). *Patterns and Meanings. Using Corpora for Language Research and Teaching*. Amsterdam: Benjamins.

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford: OUP.

Tognini-Bonelli, Elena (2001). *Corpus Linguistics at Work*. Amsterdam: Benjamins.

Collections on studies and issues in corpus linguistics:

Aarts, Jan; de Haan, Pieter, Oostdijk, Nelleke, eds. (1993). *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam: Rodopi.

Aijmer, Karin; Altenberg, Bengt, eds. (1991). *English Corpus Linguistics*. London: Longman.

Leitner, Gerhard, ed. (1992). *New Directions in English Language Corpora*. Berlin: de Gruyter.

Svartvik, Jan, ed. (1992). *Directions in Corpus Linguistics*. Berlin: de Gruyter.

Thomas, Jenny; Short, Michael, eds. (1996). *Using Corpora for Language Research*. London: Longman.

A few examples of (paper-length) corpus-based studies

Altenberg, Bengt (2002). "Modality in advanced Swedish learners' written interlanguage." In Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson, eds., *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: Benjamins, 55-76. [ICLE]

Berglund, Ylva (2000). "Utilising present-day English corpora: a case study concerning expressions of future." *ICAME Journal* 24, 25-63.

(available at <http://nora.hd.uib.no/journal.html>) [BNC, LLC, LOB, FLOB]

Gotti, Maurizio (2003c). "*Shall* and *will* in contemporary English: a comparison with past uses." In Roberta Facchinetti, Manfred Krug, Frank Palmer, eds, *Modality in Contemporary English*. Berlin / New York: Mouton de Gruyter, 267-300. [Helsinki corpus, diachronic part]

Hundt, Marianne (1998). "It is important that this study (should) be based on the analysis of parallel corpora: On the use of the mandative subjunctive in four major varieties of English." In Hans Lindquist et al., eds., *The Major Varieties of English. Papers from MAVEN '97*. Växjö: Acta Wexionensia. [LOB, Brown, FLOB, Frown, ACE, WCNZE]

Hundt, Marianne (2004). "Animacy, agentivity, and the spread of the progressive in Modern English." *English Language and Linguistics* 8 (1), 47-69. [ARCHER]

Leech, Geoffrey (2003). "Modality on the move: the English modal auxiliaries 1961-1992." In Roberta Facchinetti, Manfred Krug, Frank Palmer, eds, *Modality in Contemporary English*. Berlin / New York: Mouton de Gruyter, 223-240. [LOB, Brown, FLOB, Frown, ICE-GB, Survey of English Usage]

Skandera, Paul (2000). "Research into idioms and the International Corpus of English." In Christian Mair & Marianne Hundt, eds., *Corpus Linguistics and Linguistic Theory: Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999*. Amsterdam: Rodopi, 339-353. [ICE-East Africa]

Information on certain corpus resources:

Biber, Douglas; Edward Finegan, Dwight Atkinson (1994). "ARCHER and its challenges: compiling and exploring a representative corpus of historical English registers." In Udo Fries, Gunnel Tottie and Peter Schneider, eds., *Creating and Using English Language Corpora. Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zürich 1993*. Amsterdam & Atlanta: Rodopi, 1-13.

Granger, S., E. Dagneaux & F. Meunier, eds. (2002). *International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Greenbaum, Sidney, ed. (1996). *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon.

On corpus linguistics versus introspection:

Fillmore, Charles (1992). "'Corpus-linguistics' vs. 'computer-aided armchair linguistics'". In Jan Svartvik, ed., *Directions in Corpus Linguistics*. Berlin: de Gruyter, 35-60.

On statistics and corpus linguistics:

Oakes, Michael P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: EUP.

On corpus linguistics and linguistic theories:

Halliday, M.A.K. (1991). "Corpus studies and probabilistic grammar." In Karin Aijmer & Bengt Altenberg, eds., *English Corpus Linguistics*. London: Longman, 30-43.

Halliday, M.A.K. (1992). "Language as system and language as instance: the corpus as a theoretical construct." In Jan Svartvik, ed., *Directions in Corpus Linguistics*. Berlin: de Gruyter, 61-77.

Schönefeld, Doris (1999). "Corpus linguistics and cognitivism." *International Journal of Corpus Linguistics* 4, 137-171.

On the use of corpora for diachronic analyses:

Rissanen, Matti (1992). "The diachronic corpus as a window to the history of English." In Jan Svartvik, ed., *Directions in Corpus Linguistics*. Berlin: de Gruyter, 185-209.

On the analysis of learner corpora:

Nesselhauf, Nadja (2004). "Learner corpora and their potential for language teaching." In John Sinclair, ed., *How to Use Corpora in Language Teaching*. Amsterdam & Philadelphia: Benjamins, 125-152.