Corpus building and investigation for the Humanities:

An on-line information pack about corpus investigation techniques for the Humanities

Unit 1: Introduction

David Evans, University of Nottingham

1.1 What a corpus is

A corpus is defined here as a principled collection of naturally occurring texts which are stored on a computer to permit investigation using special software.

A corpus is principled because texts are selected for inclusion according to pre-defined research purposes. Usually texts are included on external rather than internal criteria. For example, a researcher who wants to investigate metaphors used in university lectures will attempt to collect a representative sample of lectures across a number of disciplines, rather than attempting to collect lectures that include a lot of figurative language. Most commercially available corpora are made up of samples of a particular language variety which aim to be representative of that variety. Here are some examples of some of the different types of corpora and how they represent a particular variety:

General corpora
An example of a general corpus is the British National Corpus which "… aims to represent the universe of contemporary British English [and] to capture the full range of varieties of language use." (Aston & Burnard 1998: 5). As a result of this aim the corpus is very large (containing some 100 million words) and contains a balance of texts from a wide variety of different domains of spoken and written language. Large general corpora are sometimes referred to as reference corpora because they are often used as a baseline against which judgements about the language varieties held in more specialised corpora can be made.

Specialised corpora
Specialised corpora contain texts from a particular genre or register or a specific time or context. They may contain a sample of this type of text or, if the dataset is finite and of a manageable size, for example all of Shakespeare's plays, be complete. There are numerous examples of specialised corpora; these include The Michigan Corpus of Spoken English (approximately 1.7 million words of spoken data collected from a variety of different encounters at the University of Michigan), the International Corpus of Learner English (20,000 words taken from essays of students learning English as a foreign language) and the Nottingham Health Communication Corpus (see section 5.3 for more details)

Comparable corpora
Two or more corpora constructed along similar parameters but each containing a different language or a different variety of the same language can be regarded as comparable corpora. An example of this type is the CorTec Corpus which contains examples of technical language in texts from five areas in both English and Portuguese.

Parallel corpora
These are similar to comparable corpora in that they hold two or more collections of texts in different languages. The main difference lies in the fact that they have been aligned so that the user can view all the examples of a particular search term in one language and all the translation equivalents in a second language. The Arabic English Parallel News Corpus contains 2 million words of news stories in Arabic and their English translation collected between 2001 and 2004, and is aligned at sentence level.

Historical (or diachronic) corpora
In order to study how language changes over time texts from different time periods can be assembled as a historical corpus. Two examples of this type are the Helsinki Diachronic Corpus of English Texts (containing 1.5 million words written between 700 and 1700) and the ARCHER (A Representative Corpus of Historical English Registers) corpus (1.7 million words covering the years 1650 to 1990).

Monitor corpora
A monitor corpus is one that is 'topped up' with new texts on a regular basis. This is done in such a way that "… the proportion of text types remains constant …" which means that each new version of the corpus is comparable with all previous versions. (Hunston 2002: 16). The best example of this type is the Bank of English, held at the University of Birmingham.

For more information about available corpora click here [hyperlink to 3.1]

Unlike much Chomskyan linguistics, corpus-based approaches to language study do not rely on intuitive judgements about grammaticality supported by invented examples. All the texts in a corpus are authentic examples of naturally-occurring linguistic data. As a result, the language in a corpus can be studied from both a purely linguistic point of view and from the perspective of discourse as a social and cultural phenomenon.

In order for the texts to be read by the software, they need to be stored in a machine-readable format. The most basic corpus simply consists of a set of documents in .txt format. Other information may be added to each text file, for example to indicate the source of the text, or the sex of the speakers. A corpus may be annotated in other ways e.g. part of speech tagging, but a simple, unannotated text file can be used with most corpus search software. Of course, an unformatted text file can be difficult to read, but the purpose is not in the first instance for the texts to be read linearly by a human reader. Rather the data is available to be processed by corpus search software.

The question of why you need your texts in electronic format is explored further in 2.2 [insert hyperlink to 2.2]

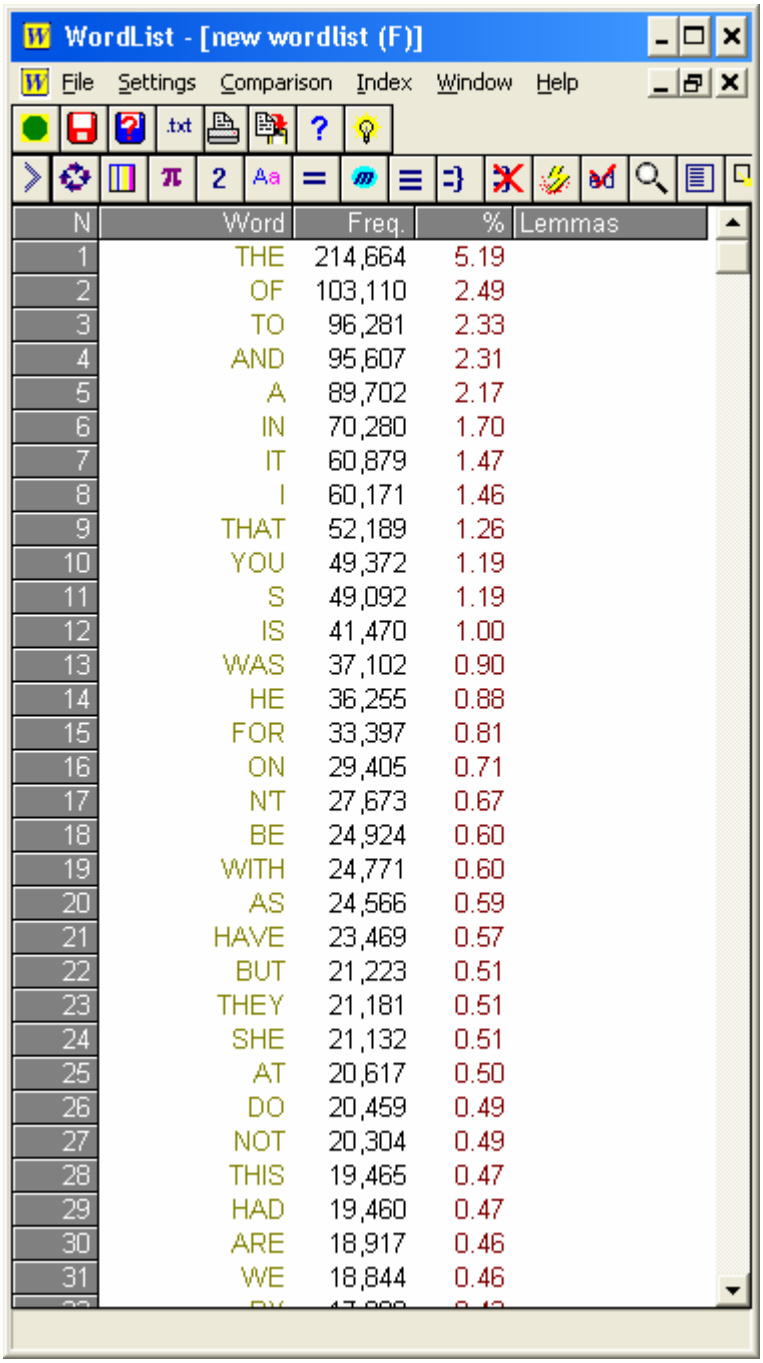
1.2 What corpus investigation software does

Corpus investigation software allows the user to process and organise large amounts of textual data relatively quickly and with a degree of accuracy that would not be possible if undertaken manually.

Corpus software performs two basic functions. It reorders the items in a corpus so that they can be observed and investigated by the user and it calculates statistical information about the data in the corpus. This reordering can be done in three basic ways: Word lists, concordances and phraseology. More details about interpreting these are given in Unit 4. The following is a brief overview of these approaches to corpus work. The examples shown use WordSmith Tools. (Scott: 1999)

Wordlists
Corpus software can break a text up according to word boundaries in order to produce word lists. These can be sorted in a variety of different ways, most commonly:

- Alphabetically, from the first character in the word (or indeed from the last character in the word, allowing the user to look at suffixes, for example, words ending with regular past suffix *–ed*)
- By frequency, as illustrated in figure 1 below.

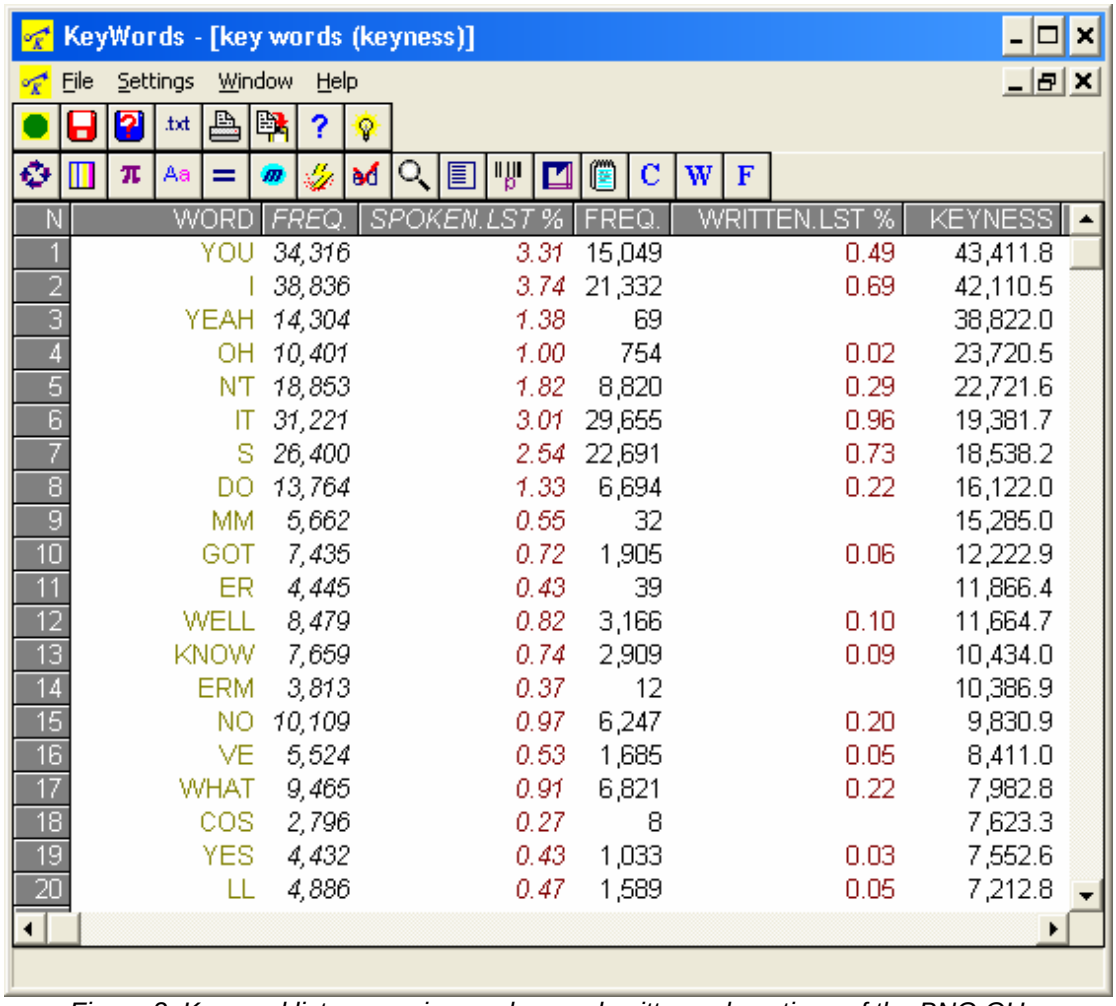


N	Word	Freq.	%	Lemmas
1	THE	214,664	5.19	
2	OF	103,110	2.49	
3	TO	96,281	2.33	
4	AND	95,607	2.31	
5	A	89,702	2.17	
6	IN	70,280	1.70	
7	IT	60,879	1.47	
8	I	60,171	1.46	
9	THAT	52,189	1.26	
10	YOU	49,372	1.19	
11	S	49,092	1.19	
12	IS	41,470	1.00	
13	WAS	37,102	0.90	
14	HE	36,255	0.88	
15	FOR	33,397	0.81	
16	ON	29,405	0.71	
17	NT	27,673	0.67	
18	BE	24,924	0.60	
19	WITH	24,771	0.60	
20	AS	24,566	0.59	
21	HAVE	23,469	0.57	
22	BUT	21,223	0.51	
23	THEY	21,181	0.51	
24	SHE	21,132	0.51	
25	AT	20,617	0.50	
26	DO	20,459	0.49	
27	NOT	20,304	0.49	
28	THIS	19,465	0.47	
29	HAD	19,460	0.47	
30	ARE	18,917	0.46	
31	WE	18,844	0.46	

Figure 1: Frequency wordlist from the BNC-OU corpus

Key word lists

The software can also take a word list, usually from a smaller, more specialised corpus and compare it with another word list from a larger, reference corpus. The resulting Keyword list prioritises the words that are most different in frequency terms in the two corpora. Figure 2 shows the key words from the spoken part of the BNC-OU corpus (a 4-million-word sample of the British National Corpus) when compared with the larger written component.



N	WORD	FREQ.	SPOKEN.LST %	FREQ.	WRITTEN.LST %	KEYNESS
1	YOU	34,316	3.31	15,049	0.49	43,411.8
2	I	38,836	3.74	21,332	0.69	42,110.5
3	YEAH	14,304	1.38	69		38,822.0
4	OH	10,401	1.00	754	0.02	23,720.5
5	NT	18,853	1.82	8,820	0.29	22,721.6
6	IT	31,221	3.01	29,655	0.96	19,381.7
7	S	26,400	2.54	22,691	0.73	18,538.2
8	DO	13,764	1.33	6,694	0.22	16,122.0
9	MM	5,662	0.55	32		15,285.0
10	GOT	7,435	0.72	1,905	0.06	12,222.9
11	ER	4,445	0.43	39		11,866.4
12	WELL	8,479	0.82	3,166	0.10	11,664.7
13	KNOW	7,659	0.74	2,909	0.09	10,434.0
14	ERM	3,813	0.37	12		10,386.9
15	NO	10,109	0.97	6,247	0.20	9,830.9
16	VE	5,524	0.53	1,685	0.05	8,411.0
17	WHAT	9,465	0.91	6,821	0.22	7,982.8
18	COS	2,796	0.27	8		7,623.3
19	YES	4,432	0.43	1,033	0.03	7,552.6
20	LL	4,886	0.47	1,589	0.05	7,212.8

Figure 2: Keyword list comparing spoken and written subsections of the BNC-OU corpus

Concordances

It is also possible to look at a search term in more in context and these are known as (KWIC) concordances, which look like this.

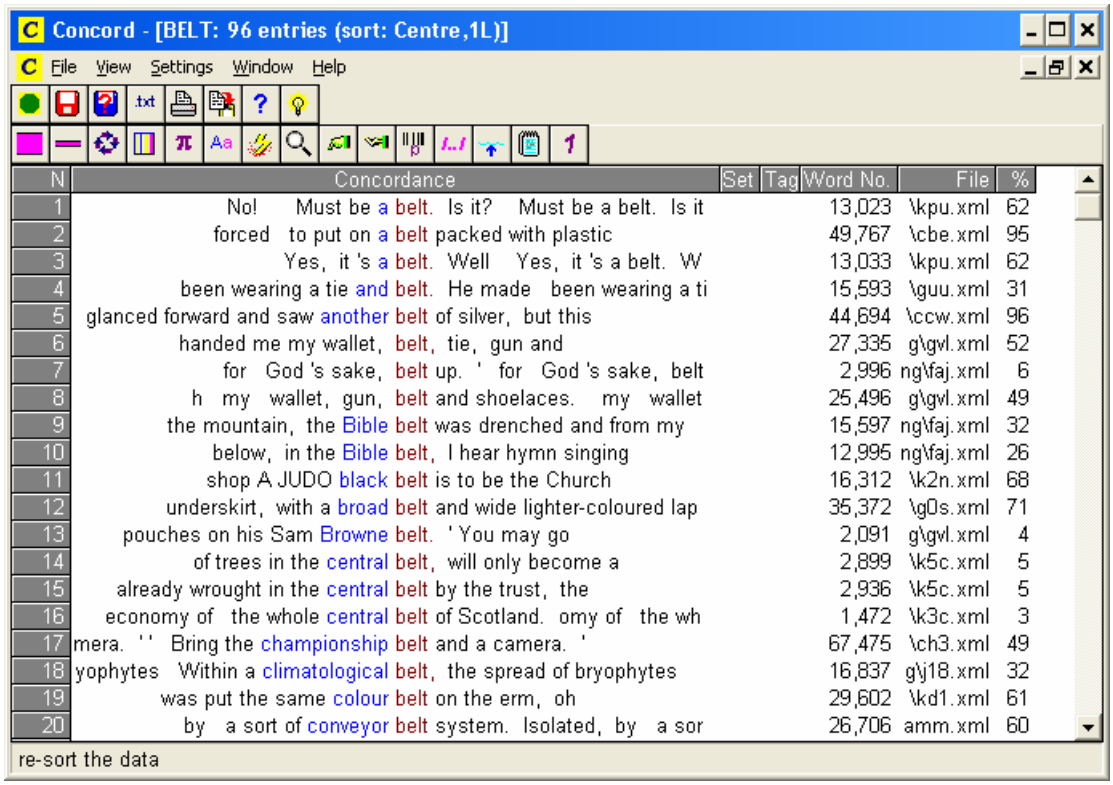


Figure 3: Concordance lines for 'belt' from the BNC-OU corpus

This facility means that the user is able to look at collocation – the partnerships that words form – which in turn has allowed corpus linguists to demonstrate that there is a much closer relationship between syntactic and lexical patterns than had previously been thought.

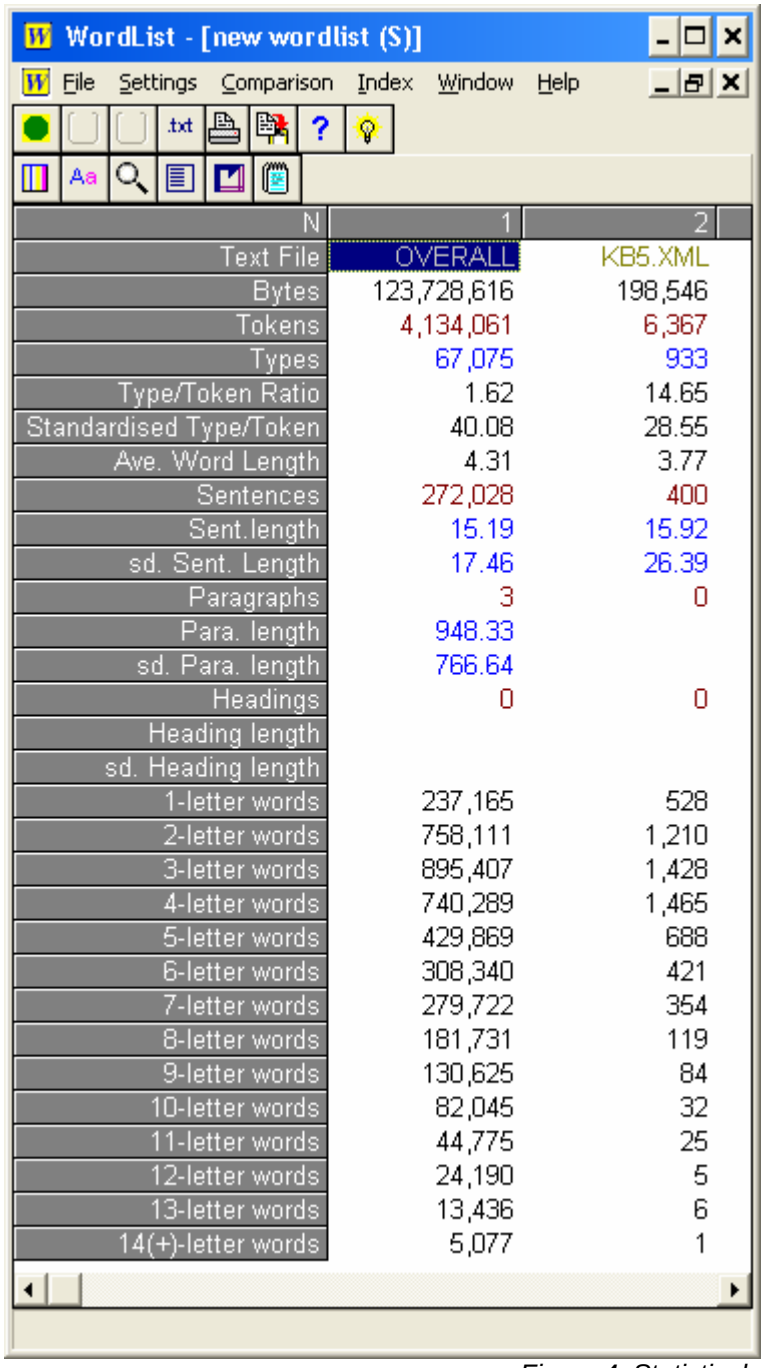
The output from a concordance can be sorted in a number of different ways. Most simply, it can be unsorted or in 'text order' i.e. the order in which the software came across the given search term. Alternatively, the words either to the left or right of the node can be sorted alphabetically. Most software allows this sorting from the word immediately to the left or right of the node, up to 5 words to the left or right. In figure 3, I was interested to see if *belt* is more frequent as a verb or a noun and if the noun is more common in its literal meaning as 'the thing that keeps your trousers up' or more metaphorically as with the examples *Bible belt* and *climatological belt*.

KWIC concordances normally look at search term in its context regardless of what that context might be. However, it is possible to instruct the software to only look at sentences or too show the original text in full. This last function can be extremely useful when trying to make sense of spoken data with its frequent false starts, repetitions and references to physical surroundings.

The search criteria are not limited to words alone. Many corpus software tools have a wildcard facility (often marked as *). This means that the search term *sort** will find examples of *sort*, *sorts*, *sorted*, *sorting*, *sortable*, etc. Alternatively, the terms *r*t* will yield all the words in the corpus beginning with 'r' and ending in 't', so from *rat* to *recalcitrant*. Most tools also allow the user to search for a sequence of words e.g. *rat catcher* or *sort of*.

Statistical data

The software can also present a wide variety of statistical data as can be seen in figure 4, with the figures for the corpus on the left, followed by figures for each individual text. At first glance this information can look a little dry but if the figures for a particular text seem disproportionate in some respect, this may prove to be a starting point for a fruitful line of investigation.



N	1	2
Text File	OVERALL	KB5.XML
Bytes	123,728,616	198,546
Tokens	4,134,061	6,367
Types	67,075	933
Type/Token Ratio	1.62	14.65
Standardised Type/Token	40.08	28.55
Ave. Word Length	4.31	3.77
Sentences	272,028	400
Sent.length	15.19	15.92
sd. Sent. Length	17.46	26.39
Paragraphs	3	0
Para. length	948.33	
sd. Para. length	766.64	
Headings	0	0
Heading length		
sd. Heading length		
1-letter words	237,165	528
2-letter words	758,111	1,210
3-letter words	895,407	1,428
4-letter words	740,289	1,465
5-letter words	429,869	688
6-letter words	308,340	421
7-letter words	279,722	354
8-letter words	181,731	119
9-letter words	130,625	84
10-letter words	82,045	32
11-letter words	44,775	25
12-letter words	24,190	5
13-letter words	13,436	6
14(+)-letter words	5,077	1

Figure 4: Statistical data from the BNC-OU corpus

1.3 The sorts of things that a corpus can help you with

"It is no exaggeration to say that corpora, and the study of corpora, have revolutionised the study of language, over the last few decades." (Hunston 2002: 1) The following section outlines some of the areas where corpora have had an impact. The intention is to help you to see whether corpus analysis techniques may be useful to you in your research.

Even if you have never used a corpus before, it is increasingly likely that you have used dictionaries and grammar books which were written using information derived from corpora as their bases, especially if English is not your first language. The following section looks at how corpora have been used to enhance understanding in three areas: translation, stylistics and language and ideology.

Translation
The parallel and comparable corpora that were mentioned in 1.1 can be used for both practical and theoretical translation studies. On a practical level, a parallel corpus can be used by a translator to look at a number of alternatives for a particular term and aid in the solution of a translation problem. A parallel corpus is a richer resource than a bilingual dictionary as it allows the user to see the search term with more of the co-text and with a broader range of contexts and collocates. This in turn shows the translator a wide range of possible renderings: from the 'zero' option, where something has been missed out by the translator, possibly for pragmatic reasons, to a phrase which differs a great deal in terms of lexical equivalence but retains the semantic content of the original.

On a more theoretical level it is possible to compare a corpus of texts translated into a language with those originally written in that language. Studies of this nature have shown how original and translated texts differ in particular ways. For example, Laviosa (1997: 315 see Hunston 2002: 127) has shown how translations are often less lexically varied than their 'original' equivalents and McEnery et al (2006: 93) demonstrate that "… the frequency of aspect markers in Chinese translations is significantly lower than that in the comparable L1 Chinese data." This information may be useful for those who study how translators work, or who are involved in the training of translators to help their students to avoid 'translationese' creeping into their work.

Stylistics
There are a number of ways in which corpus-based approaches can contribute to the study of not just literary works but 'literariness' in general. The statistical analysis of literary texts, known as Stylometrics, has been used to establish authorship of contested texts. As has been mentioned before, a smaller corpus of literary texts can be compared with a reference corpus to investigate literary 'devices' to see how they vary from more 'everyday' varieties of English.

Louw (1997: 245) demonstrates how students can confirm their intuitions about literary texts using corpus data. In one example, students investigated the term *wielding a* in order to confirm that the use of this term in the line 'And crawling sideburns, wielding a guitar' from the poem *Elvis Presley* by Thom Gunn was being used ironically. As expected they found that *wielding a* is most frequently used with some kind of weapon. What was unexpected was the very high frequency with which the term is used ironically, prompting one student to comment that it may soon lose its power as a writer's device.

Language and ideology
There is an increasing interest in using corpora to investigate the ideological stance of writers and speakers in texts. Frequently occurring patterns allow the observer to make deductions about what a group or society sees as valuable or important. Information about collocation means that new concepts and the range of associations of a word can be monitored. Stubbs (1996: 195) argues that if a collocation becomes more common in the language then it is more likely to become fixed in the minds of speakers and therefore, more difficult to challenge. As we saw with stylistics, semantic prosody, the semantic associations of a word or phrase, can be used to carry covert messages.

Studies in this area have covered a wide variety of areas such as sexism and racism in media discourse, Euroscepticism, political correctness and the difference in rhetorical styles of Bush and Blair in relation to the war in Iraq (see McEnery et al 2006: 108-113). Hunston (2002: 121) points out some of the assumptions that such studies can be based on. O'Halloran and Coffin (2004) argue that using corpora can actually help the researcher to avoid over- and under-interpretation when working with texts. While caution should be exercised regarding the verifiability of claims about ideology found in corpora, they remain valuable resource in such studies.

1.4 What you need to do corpus work

You can actually get started on some corpus work straight away, if you have internet access. There are corpora that you can browse (although not always in full) online. Examples include:
MICASE [http://www.lsa.umich.edu/eli/micase/index.htm]
BNC [http://www.natcorp.ox.ac.uk/]
Business Letter Corpus [http://ysomeya.hp.infoseek.co.jp/]

Or you can make concordances from the World Wide Web using the tools that can be found at these sites:
WebCorp [http://www.webcorp.org.uk/]
WebCONC [http://www.niederlandistik.fu-berlin.de/cgi-bin/web-conc.cgi?art=google&sprache=en]

If you intend to install some corpus investigation software onto a computer then the more RAM and the faster the processor, the easier the computer will be able to handle the tasks you might ask of it. Much of the software that has been developed thus far has been written for use with Windows operating systems. For Mac users the advent of Apple Boot Camp now makes it possible to run such software on a Mac.

If you want to look into some of the software that's available go here [insert hyperlink to 3.2]. If you are looking to use some existing corpora then click here [insert hyperlink to 3.1]. If you are thinking of designing and building your own corpus then go to Unit 2 [insert hyperlink to Unit 2].

References and further reading

Aston, G. & Burnard, L. (1998) *The BNC Handbook* Edinburgh: Edinburgh University Press

Coffin, C. & O'Halloran, K. (2004) Checking Overinterpretation and Underinterpretation: Help from Corpora in Critical Linguistics in Coffin, C. Hewings, A. & O'Halloran, K. (eds.) *Applying English Grammar* London: Arnold

Hunston, S. (2002) *Corpora in Applied Linguistics* Cambridge: Cambridge University Press

Kennedy, G. (1998) *An Introduction to Corpus Linguistics* Harlow: Longman

Louw, B. (1997) The Role of Corpora in Critical Literary Appreciation in Wichmann, A., Fligelston, S., McEnery, T. & Knowles, G. (eds) *Teaching and Language Corpora* Harlow: Longman

McEnery, T, & Wilson, A. (1996) *Corpus Linguistics* Edinburgh: Edinburgh University Press

McEnery, T., Xiao, R. & Tono, Y. (2006) *Corpus-Based Language Studies* Abingdon: Routledge

Meyer, C. (2002) *English Corpus Linguistics* Cambridge: Cambridge University Press

Scott, M. (1999) *WordSmith Tools* Oxford: Oxford University Press

Stubbs, M. (1996) *Text and Corpus Analysis* Oxford: Blackwell