

Published in: Gerbig, Andrea & Anja Müller-Wood (eds.). 2006. *How Globalization Affects the Teaching of English: Studying Culture Through Texts*. Lampeter: E. Mellen Press. 81-109.

Where the Computer Meets Language, Literature, and Pedagogy: Corpus Analysis in English Studies

Ute Römer

University of Hanover

1 Introduction: Corpora and English Studies

In the field of linguistics, electronic corpora (i.e. large systematic collections of spoken and/or written text stored on the computer) and corpus-analytic tools have been widely used for more than forty years now. For very good reasons, linguists all over the world draw on corpora in language analysis and description, and it is part of the daily working routine of a steadily growing number of linguistic researchers to run software programs in order to get access to language patterns, collocations, and word frequencies, or simply to retrieve examples of a particular lexical item in use. Since they enable, to a high degree, objective views on the language, these resources can help us discover what language is really like, which means that we no longer have to speculate and rely on our (or other individuals') intuition as the sole source of data.

The aim of this essay is to show that corpora and corpus-analytic tools can not only be extremely valuable in linguistics but also in language pedagogy and in literary studies. It will be explored in what ways some basic techniques of corpus analysis can provide insights into language and literature – insights that intuitive approaches to the same objects of study may fail to create. Hence, one of the questions I will address in the following is, "Can corpus analysis help to improve interdisciplinarity in English Studies?" In other words, referring to the title of the present volume, it will be asked, "Can corpora contribute to a reconciliation of

literature, linguistics, and pedagogy?" These three branches of English as a subject are usually studied separately and, more often than not, isolated in research and teaching. In addressing these questions, I mainly aim to open up new perspectives on working with texts to the non-(corpus)linguist, but I also wish to put some emphasis on the student perspective. As the editors of this volume rightly note in their introduction, we as researchers and teachers of English linguistics, didactics, or literary and cultural studies often fail to explain to our students what and where the common ground is between these disciplines. In research and teaching we strictly separate what, in fact, belongs together and should form an integrated whole. It is hence not surprising that students find it difficult to spot the existing, but often well-hidden, relations between the branches of the subject. In the present paper I shall argue that the profitable use of one particular tool across these separated branches can mean one important step towards their reconciliation. The corpus and everything that is directly related to its use is put forward as a reconciliatory means and as a key tool in the study of English in general.

Before the universality of the application of corpora and corpus-analytic methods in English Studies will be illustrated by means of a few case studies from the individual subfields, however, I shall provide an overview of some central steps in the corpus-analytic exploration of text.

2 Corpus Analysis: Some Basic Techniques

As mentioned above, a corpus is essentially a computerised collection of text used for linguistic analysis, no more, no less. With the help of available software programs, this text collection on the computer can be accessed in a number of different ways (cf. Barlow 2004).

In corpus linguistics access to the texts stored in a corpus is enabled by so-called concordance programs or concordancers, i.e. software packages which provide a range of functions to analyse a collection of texts, the most important

one being the concordance function (exemplified below). Without such concordance programs, corpora would be of no use other than being repositories of texts – texts that could then be read on screen in the normal linear fashion. But what exactly does this corpus analysis software do? In Hunston and Francis' (2000: 15) terms, it "selects, sorts, matches, counts and calculates". With reference to one particular software package, *WordSmith Tools* (Scott 1999), I will illustrate how this is done, and present some basic techniques in accessing and analysing corpora.¹ Following Barlow (2004: 205), I will regard corpus or text analysis as corpus or text transformation. I will try to show in what ways different types of transformation can draw attention to different aspects of a text.

The corpus that will be used to illustrate these basic analytic (or transformational) procedures consists of only one text: T.S. Eliot's poem "The Love Song of J. Alfred Prufrock" (henceforth "Prufrock"). The first ten lines of the poem are displayed in figure 1. Admittedly, this is a tiny corpus and a very specialised one too, but it will fully serve our illustratory purpose.

Let us go then, you and I,
 When the evening is spread out against the sky
 Like a patient etherized upon a table;
 Let us go, through certain half-deserted streets,
 The muttering retreats
 Of restless nights in one-night cheap hotels
 And sawdust restaurants with oyster-shells:
 Streets that follow like a tedious argument
 Of insidious intent
 To lead you to an overwhelming question...

Figure 1: Lines 1-10 of T.S. Eliot's poem "The Love Song of J. Alfred Prufrock"

¹ Another well-known concordance package used in corpus analysis is *MonoConc Pro* (Barlow 2000). For a comparison of features of *MonoConc Pro* and *WordSmith Tools*, readers are referred to a review published in *Language Learning & Technology* (Reppen 2001), available online at <http://llt.msu.edu/vol5num3/review4/default.html> (consulted: Oct. 3rd 2004).

I downloaded "Prufrock" from the web, more precisely from the Project Gutenberg website,² and saved it on the harddrive of my computer in text file format as 'prufrock.txt'. This text file could then be analysed with the software package *WordSmith Tools*. The analytic procedure was carried out in five steps.

Step 1: Word Listing and Counting – Tearing the Text Apart

As Barlow (2004: 207) notes, "[p]robably the most radical transformation of a text used in linguistic analysis is to, in effect, rip it apart to produce a wordlist." This is what I did with Eliot's dramatic monologue in the first corpus-analytic step. Using *WordSmith's* 'WordList' tool, the statistics of the poem were determined and a frequency wordlist was created (cf. figure 2).

We can see in the right-hand screenshot in figure 2 that our mini-corpus consists of 405 different words (types) which altogether occur 1,063 times (i.e. the corpus contains 1,063 tokens). We also find further statistical information on average word and sentence lengths and on the numbers of 1-, 2-, 3-, etc. letter words in the poem. Perhaps more interesting than statistics, however, is the frequency ranking of the words found to occur in the corpus. The left-hand part of figure 2 shows that, with 74 occurrences, *the* is the most frequent word in "Prufrock" and that the word makes up 6.96% of the whole text. Also particularly common in the text are *and, I, a, to, have, that, and of*.

² Project Gutenberg produces free electronic texts on the internet. More than 12,000 e-books are currently available for download from <http://www.gutenberg.net/> (consulted: Oct. 3rd 2004). A large number of literary texts is also available from the Oxford Text Archive at <http://ota.ahds.ac.uk/> (consulted: Oct. 3rd 2004).

N	Word	Freq.	%	N	
1	THE	74	6.96	1	Text File
2	AND	57	5.36		Bytes
3	I	41	3.86		Tokens
4	A	34	3.20		Types
5	TO	23	2.16		Type/Token Ratio
6	HAVE	22	2.07		Standardised Type/Token
7	THAT	21	1.98		Ave. Word Length
8	OF	20	1.88		Sentences
9	IT	15	1.41		Sent. length
10	ALL	14	1.32		sd. Sent. Length
11	IN	13	1.22		Paragraphs
12	MY	11	1.03		Para. length
13	TIME	11	1.03		sd. Para. length
14	IS	10	0.94		Headings
15	WILL	10	0.94		Heading length
16	SHOULD	9	0.85		sd. Heading length
17	YOU	9	0.85		1-letter words
18	AFTER	8	0.75		2-letter words
19	FOR	8	0.75		3-letter words
20	AT	7	0.66		4-letter words

Figure 2: Frequency-ordered wordlist based on T.S. Eliot's "Prufrock" (left) and statistics of the poem (right)

In the analysis of such a wordlist we now have two options. One option is to analyse the wordlist manually, i.e. to simply scroll down the list, in order to see which words, in particular which content words (such as nouns, verbs, and adjectives), are especially common, and hence to trace some peculiarities of the underlying text. The second option is to compare this wordlist, which is based on a very specialised corpus (the poem "Prufrock" by T.S. Eliot), with another wordlist derived from a more general so-called reference corpus, i.e. a corpus that is considerably bigger and to a high degree representative of written English in general. In the *WordSmith Tools* program suite this can be done by means of the 'compare 2 wordlists' command in the WordList tool. As a result, we get a list of keywords in the smaller corpus, ordered by their 'keyness' in the text as compared to the selected reference corpus, which was in our case the 90 million-word written part of the British National Corpus (BNC_written).

Figure 3 displays the thus retrieved keywordlist of "Prufrock" with the most outstanding words topping the list. Even though they are very frequent in the poem (cf. figure 2), words like *the*, *a*, *to*, and *of* do not appear in this list. This is because these words are generally extremely common in written English and not key items of Eliot's poem. The items listed in figure 3 in order of keyness (*I*, *panes*, *have*, *dare*, *presume*, *upon*, etc.) are all more frequent in "Prufrock" than we would expect them to be on the basis of general language use. Knowing which words in a text, or in a collection of texts, are unexpectedly or unusually frequent can help us determine what the text is concerned with. The compilation of a keywordlist like the one given in figure 3 is hence a very useful step in the characterisation of any text or text collection.

N	WORD	FREQ.	PRUFROCK.TXT %	FREQ.	BNCWRIT.LST %	KEYNESS
1	I	41	3.86	603,990	0.64	80.0
2	PANES	3	0.28	186		37.6
3	HAVE	22	2.07	395,993	0.42	35.4
4	DARE	4	0.38	1,868		34.0
5	PRESUME	3	0.28	397		33.0
6	UPON	7	0.66	22,823	0.02	32.8
7	KNOWN	7	0.66	26,917	0.03	30.6
8	WORTH	5	0.47	11,190	0.01	27.1
9	STREETS	4	0.38	4,758		26.6
10	RUBS	2	0.19	91		26.3
11	AM	6	0.56	24,424	0.03	25.6
12	MY	11	1.03	129,119	0.14	25.5
13	HAIR	5	0.47	13,631	0.01	25.2
14	AND	57	5.36	2,431,895	2.58	25.1
15	MICHELANGELO	2	0.19	125		25.0

Figure 3: Keywordlist based on T.S. Eliot's "Prufrock" and BNC_written as reference corpus

Step 2: Tracing the Repeated Occurrence of an Item in a Text – Examining Dispersion Plots

If we do not only want to know *what* the most common (or the most unusually common) words in our corpus are but also *where* they occur in the text, we can

make use of the 'dispersion plot' function in *WordSmith Tools* (comparable to the 'distribution of hits' function on *MonoConc Pro*). This function serves to visualise the distribution of a selected item across a text.

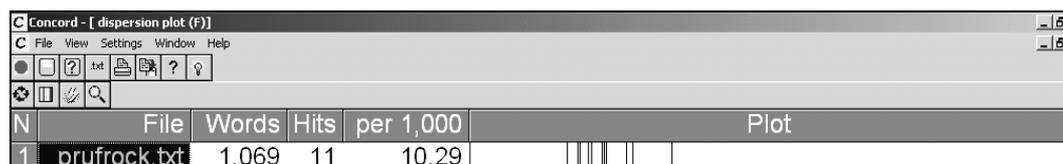


Figure 4: Dispersion plot of *time* in T.S. Eliot's "Prufrock"

Figure 4 shows the dispersion plot of the most frequent noun, the noun *time*, in "Prufrock".³ Each of the eleven occurrences of the word in the text is represented by a thin black line in the 'plot' window. The first line in the plot column (on the left) indicates the beginning, the last line on the right indicates the end of the text. As we can see, *time* clusters in the first third of "Prufrock". There are no instances of the word at the very beginning of the poem and none in the second half of it. We can therefore infer that *time* only features in one or two stanzas in the first third of "Prufrock" and that it is not a theme traceable throughout the entire poem.

Step 3: Compiling a Concordance – Putting Words Back into Context

In the third analytic step we are moving towards the centre of each kind of corpus-linguistic activity: the concordance. Having torn the text apart in step 1, we will now reverse this process and provide a contextualised view of those items in our corpus in which we are most interested (maybe some of the particularly common or key words).

As fittingly stated by Barnbrook (1996: 65), "[t]he concordance provides a simple way of placing each word back in its original context, so that the details of its use and behaviour can be properly examined." A concordance of

³ *Time* is in fact not only the most frequent noun in "Prufrock" but also in English in general. Hence the item does not appear in the keywordlist displayed in figure 3.

the word *time* based on our "Prufrock" corpus is given in figure 5. Generally, in such a concordance all occurrences (tokens) of a word are shown in the context of a concordance line, usually displayed in the so-called 'KWIC' (key word in context) format, with the searchword (here *time*) in the middle of the screen and some context on the left and on the right of it. In our case the concordance shows all eleven instances of *time* in "Prufrock" in random order.

The screenshot shows a window titled "Concord - [TIME: 11 entries (sort: Centre)]". The window contains a menu bar (File, View, Settings, Window, Help) and a toolbar with various icons. Below the toolbar is a table with the following content:

N	Concordance
1	our plate; Time for you and time for me, And time yet for a hundred indecisions, And for a
2	p a question on your plate; Time for you and time for me, And time yet for a hundred inde
3	s That lift and drop a question on your plate; Time for you and time for me, And time yet f
4	e To wonder, "Do I dare?" and, "Do I dare?" Time to turn back and descend the stair, Wit
5	ng of Michelangelo. And indeed there will be time To wonder, "Do I dare?" and, "Do I dar
6	re Disturb the universe? In a minute there is time For decisions and revisions which a mi
7	down-panes; There will be time, there will be time To prepare a face to meet the faces th
8	se, and fell asleep. And indeed there will be time For the yellow smoke that slides along
9	There will be time to murder and create, And time for all the works and days of hands Th
10	meet the faces that you meet; There will be time to murder and create, And time for all t
11	back upon the window-panes; There will be time, there will be time To prepare a face to

Figure 5: Unsorted concordance of *time* in T.S. Eliot's "Prufrock"

One major advantage of this concordance function is that we do not have to read through the whole text if we want to see how a particular word is used. All we need is one mouse-click and the required textual evidence is displayed on the screen right in front of us. However, access to the usage patterns in such a concordance display can be further optimised, as the next step in our analysis shows.

Step 4: Sorting the Context in a Concordance – Uncovering Patterns

Concordance programs usually offer the possibility of sorting or re-sorting the concordance, which means that the order of the displayed concordance lines is rearranged according to certain predefined sorting criteria. In the analysis of

"Prufrock" I decided to sort the context in the *time* concordance alphabetically to the left. The result of this sorting step can be seen in figure 6.

N	Concordance
1	There will be time to murder and create, And time for all the works and days of hands Tha
2	our plate; Time for you and time for me, And time yet for a hundred indecisions, And for a
3	p a question on your plate; Time for you and time for me, And time yet for a hundred inde
4	ng of Michelangelo. And indeed there will be time To wonder, "Do I dare?" and, "Do I dar
5	se, and fell asleep. And indeed there will be time For the yellow smoke that slides along
6	dow-panes; There will be time, there will be time To prepare a face to meet the faces th
7	meet the faces that you meet; There will be time to murder and create, And time for all t
8	back upon the window-panes; There will be time, there will be time To prepare a face to
9	e To wonder, "Do I dare?" and, "Do I dare?" Time to turn back and descend the stair, Wit
10	re Disturb the universe? In a minute there is time For decisions and revisions which a mi
11	s That lift and drop a question on your plate; Time for you and time for me, And time yet f

Figure 6: Left-sorted concordance of *time* in T.S. Eliot's "Prufrock"

In contrast to the unsorted concordance in figure 5, the sorted display in figure 6 enables us to spot at a glance repeated word combinations in the text, e.g. "and time" (three times) or "there will be time" (five times). Sorting to the right of the searchword, which is of course also possible, would highlight different patterns, such as "time for" (five times) and "time to" (four times). Access to language patterns is absolutely central in all sorts of text analysis, since we can only fully understand what is expressed by a word when we see it in combination with other words. In Wittgenstein's terms, "the meaning of a word is its use in the language" (*Philosophical Investigations*, §43). Granting us direct access to patterns of language use, sorted concordances can thus help us determine how meanings are made. The fifth and final corpus-analytic step that I would like to demonstrate will further focus on the important issue of repeated combinations and co-occurrences of words.

Step 5: Examining the Context of a Word – Looking for Collocations

There are two functions in the *WordSmith* 'Concord' tool (the tool I used to create the *time* concordance) that serve to highlight co-occurrence patterns in texts: the 'show collocates' and the 'clusters' function. By collocates we mean words that occur near the searchword, usually in a contextual span of four or five words to the left and to the right of it.



The screenshot shows a window titled "Concord - [collocates (total)]". The window contains a menu bar (File, View, Settings, Window, Help) and a toolbar with various icons. Below the toolbar is a table with the following data:

N	WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	* R1	R2	R3	R4	R5	
1	TIME	17	3	3	0	3	0	0	0	11	0	0	0	2	1
2	AND	15	8	7	3	1	1	0	3	0	0	0	3	2	2
3	FOR	9	2	7	0	0	2	0	0	0	3	3	0	0	1
4	BE	7	6	1	1	0	0	0	5	0	0	0	0	1	0
5	THERE	7	6	1	0	0	5	1	0	0	0	1	0	0	0
6	WILL	6	5	1	0	0	0	5	0	0	0	0	1	0	0
7	TO	5	1	4	1	0	0	0	0	0	2	2	0	0	0
8	A	4	2	2	1	1	0	0	0	0	0	0	1	1	0
9	YOU	3	2	1	1	0	0	1	0	0	0	1	0	0	0

Figure 7: Collocates of *time* in T.S. Eliot's "Prufrock"

The activation of the 'show collocates' function, based on the concordance of *time*, retrieves the collocation plot given in figure 7. The different columns list the most common collocates of *time* and their numbers of occurrence on either side of the searchword, split up according to (in our case) five positions to the left (L1 to L5) and five positions to the right (R1 to R5). This display of numbers tells us, for instance, that of the six co-occurrences of *time* and *will*, *will* is found five times in L2 position, that is two places on the left of *time* as in "there will be time". The 'clusters' function facilitates access to such collocation patterns in that it automatically extracts all recurrent word combinations from a concordance that matches a certain predefined cluster span. I defined a cluster span of '3' to extract all three-word clusters that occur at least twice in the *time* concordance (see figure 8).

N	cluster	Freq.
1	there will be	6
2	will be time	6
3	and time for	4
4	be time to	4
5	and indeed there	2
6	and time yet	2
7	for you and	2
8	indeed there will	2
9	murder and create	2
10	time for me	2
11	time for you	2
12	to murder and	2
13	you and time	2

Figure 8: Repeatedly occurring three-item clusters based on the concordance of *time* in T.S. Eliot's "Prufrock"

Also important in this context of investigating typical collocations and co-occurrence patterns of words is the notion of 'semantic prosody'. When we look at the items on the right-hand side of *time* in figure 5, we note a number of words and phrases with rather negative connotations. There is time "to murder", "for a hundred indecisions", "for the yellow smoke", or time "to turn back and descend the stair". These words and phrases all contribute to the meanings of *time* in "Prufrock". The, in this case, negatively connotated meaning they create is called semantic prosody. Sinclair (2003: 117) refers to semantic prosodies as the "[h]idden meanings" of words, which are "essential for the understanding of language text" (ibid.) and which can be uncovered by means of corpus-analytic techniques (cf. also Bublitz 1996, Esser 1999, Louw 1993, and Partington 2004).

To sum up, the central steps in the analysis and thus in the transformation of a corpus are: counting and listing words, tracing their distribution across texts, compiling concordances, sorting concordances, and searching for collocations and patterns (including the determination of semantic prosodies).

3 The Use of Corpus-Analytic Techniques in English Studies

3.1 Corpus Analysis and Linguistics

Probably the most obvious field of application of the outlined steps in corpus and text analysis is linguistics. Linguists study language, its production, its structure, and its use, and they need a certain amount of language data on which to base their research. Corpora therefore present welcome tools in all traditionally distinguished fields of linguistics, be it phonology, morphology, lexicology, syntax, semantics, pragmatics, historical linguistics, sociolinguistics, or psycholinguistics.⁴ To provide a rough idea of what kinds of insights the use of corpora and corpus-analytic techniques may provide in linguistics, I shall now briefly sketch a research project I carried out on English modal verbs (for details cf. Klages & Römer 2002 and Römer 2004a).

The project centres on the use of the central English modals (*can*, *could*, *may*, *might*, *will*, *would*, *shall*, *should*, *must*, and *ought to*) in spoken British English. Two key questions I dealt with in this project were, "How are different modal verbs distributed?" and "Which meanings do they express?" Of major interest in tackling both questions were frequencies of occurrence of the modals and their meanings in actual language use. The first step in my analysis thus involved some word counting. With the help of a concordancer I determined the frequencies of the above-listed modal verbs in the spoken component of the British National Corpus (including their negative forms and excluding instances in which the items do not function as modals, e.g. *can* in "I'll put a can of that juice in") and established the frequency ranking displayed in figure 9. We can see in this figure that in spoken British English *will*, *would*, and *can* are by far most

⁴ For overviews on the use of corpora in various linguistic sub branches the reader is referred to the following introductory corpus-linguistic textbooks: Biber, Conrad & Reppen 1998, Hunston 2002, McEnery & Wilson²2001, Meyer 2002, and Partington 1998.

common among all modal verbs and that verbs like *may*, *must*, or *shall* are comparatively infrequent.

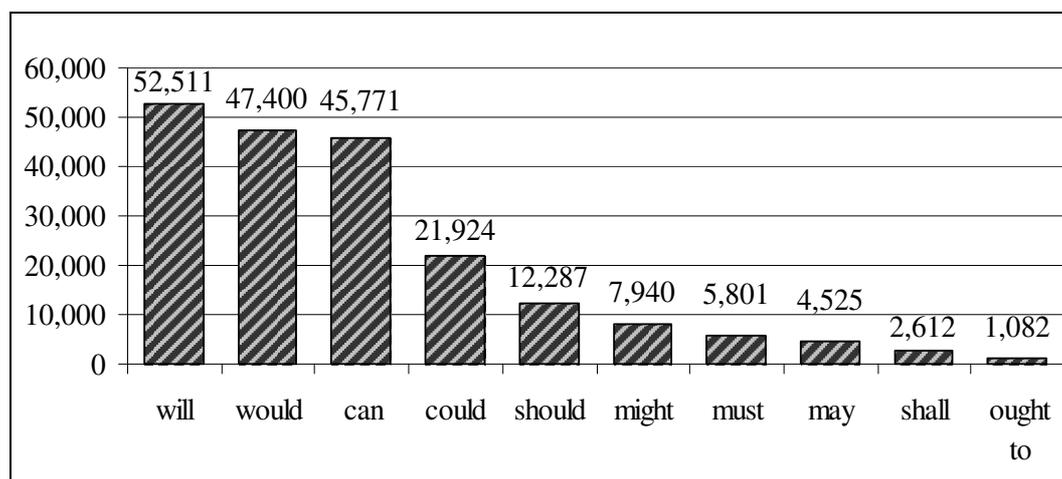


Figure 9: Overall frequencies of the central modal verbs in BNC_spoken

As each modal can express different meanings, depending on the immediate lexical context in which it occurs, I considered it important to move one step further in the analysis, from overall frequencies to the distribution of these different modal meanings. Again, a concordance program was consulted and concordances were compiled in order to retrieve a large number of examples of modal verbs in context. An extended context window in the concordances (of 200 characters per line) enabled me to attach a meaning label to almost every example⁵ and so to determine the shares of different functions for each modal. Figure 10 visualises the results of this meaning attribution in the case of *may*, a modal that expresses possibility and permission, however not with equal shares.

As figure 10 shows, instances of *may* in which a speaker grants or asks for permission (as in "May I ask a question?") are much less common in spoken British English than occurrences of 'possibility' *may* (e.g. "It may well be that their

⁵ Mainly due to the often fragmentary nature of spoken English, there were a couple of cases in the datasets which could not be clearly attributed to either of the defined meaning categories. These examples were labelled 'unclear'.

expectation is wrong."). Besides, the manual analysis of the context in the sorted concordance of *may* (and the other modals) gave access to co-occurrence patterns and typical collocations. Among the things I observed were a rather high share of if-clauses in the *may* concordance set, and a repeated occurrence of the cluster "if I may".

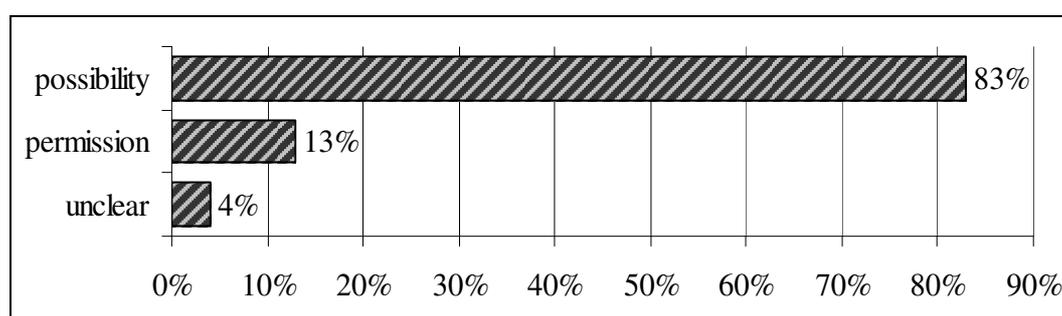


Figure 10: The distribution of different meanings of *may* in BNC_spoken

These are just some of the findings I was able to retrieve using corpus-analytic techniques in my approach to an object of linguistic study. Such empirical findings are not only relevant for a systematic stock-taking of the structure and use of a language, they can also have important implications for language learning and teaching – a field that forms another part of the subject 'English' as it is taught and studied in a large number of English departments across Europe. Before further investigating the impact of corpora and corpus analysis on language pedagogy in section 3.3 of this essay, I will first turn to the application of corpus-analytic procedures in the study of literature.

3.2 Corpus Analysis and Literature

The aim of this section is *not* to give an introduction to or provide an overview of work done in the field of stylistics. Rather than asking questions like "What is style?" and "How can it be captured in linguistic terms?", I will address the question "How can corpus analysis help to shed light on aspects about an author's

style?"⁶ Taking for granted that linguistic approaches to literary language can be fruitful and illuminating, I will try to exemplify the specific surplus value of corpus-analytic techniques in the study of literature and show how the steps described in section 2 of this paper can be applied to a corpus of literary works. I wish to stress here that corpus analysis is not supposed to replace, but to complement, traditional approaches in literary analysis.

The corpus in question consists of the six major novels written by the early nineteenth century British novelist Jane Austen (the Jane Austen Corpus, henceforth JAC) and was compiled by Katrin Oltmann as part of a university research project in corpus linguistics.⁷ In the following I will largely draw upon Oltmann (2004) and present selected results from her study on irony in Austen's writing, complemented by findings from some additional queries based on JAC and a larger corpus of eighteenth and nineteenth century English novels (the Novels Reference Corpus, henceforth NRC).^{8,9}

Jane Austen and her Major Concerns – Keywords in JAC

As pointed out above (cf. section 2), the comparison of a frequency wordlist from a small and specialised corpus with a frequency wordlist based on a larger reference corpus can highlight items in the smaller corpus ('keywords') that are unexpectedly common and that are likely to characterise the texts in this corpus. The type of quick analysis that I carried out on keywords in Jane Austen's novels could be used as a pre-reading activity or as a preliminary step in our, or in our

⁶ For further studies which use corpus approaches in the analysis of literary works, the reader is referred to Krishnamurthy's (1995) article on the poem "Spring" by Philip Larkin, Stubbs' (2001) chapter on "Eveline" by James Joyce (from *Dubliners*), his analysis of Conrad's *Heart of Darkness* (this volume), and Sinclair's (forthcoming) paper on Alexander Pope's *An Essay on Man*.

⁷ JAC has an overall size of roughly 650,000 words (tokens) and includes the novels *Emma*, *Mansfield Park*, *Northanger Abbey*, *Persuasion*, *Pride and Prejudice*, and *Sense and Sensibility*. The texts were downloaded from the Project Gutenberg online archive (see <http://www.gutenberg.net/>; consulted: Oct. 3rd 2004).

⁸ I would like to thank Katrin Oltmann for granting me permission to use JAC and NRC for this study.

⁹ The texts in the Novels Reference Corpus (NRC) were also taken from Project Gutenberg. The NRC has a size of roughly 4.4 million words.

students', interpretation of any literary work (on this issue see also Kettemann & Marko 2004).

N	WORD	FREQ.	JAC.LST %	FREQ.	NRC.LST %	KEYNESS
1	FANNY	858	0.13	37		3,217.2
2	EMMA	785	0.12	1		3,200.9
3	HER	12,028	1.86	43,949	1.01	3,191.6
4	ELINOR	623	0.10	1		2,537.6
5	SHE	9,121	1.41	33,116	0.76	2,457.0
6	CRAWFORD	493	0.08	0		2,019.5
7	ELIZABETH	628	0.10	120		1,947.0
8	MARIANNE	492	0.08	27		1,810.7
9	BE	7,286	1.13	27,358	0.63	1,773.9
10	DARCY	374	0.06	0		1,532.0
11	HARRIET	419	0.06	31		1,499.2
12	EDMUND	365	0.06	0		1,495.1
13	KNIGHTLEY	356	0.06	0		1,458.3
14	VERY	3,369	0.52	10,195	0.23	1,413.1
15	MISS	1,739	0.27	3,552	0.08	1,404.9
16	MRS	2,164	0.33	5,333	0.12	1,328.5
17	ELTON	320	0.05	0		1,310.8
18	NOT	7,663	1.18	32,544	0.75	1,219.2
19	WESTON	389	0.06	79		1,190.3
20	BENNET	294	0.05	4		1,162.9

Figure 11: Keywords in JAC (items 1-20), sorted according to keyness

Figure 11 displays the first 20 items in a keywordlist based on JAC, sorted by keyness values, with NRC used as reference corpus. A number of mainly female first names (e.g. *Fanny*, *Emma*, and *Elinor*) top this list, which – given the composition of the underlying corpus – is hardly surprising. In fact, if we scroll further down the JAC keywordlist, we note that the majority of the items listed are names of novel characters or places. This observation tells us something about *who* Austen's heroines and heroes are and how often they are referred to in the texts, but it does not reveal much about *what* her novels are actually concerned with. In a further analytic step I have, therefore, extracted all keywords that are not names and listed the first 40 items in table 1 (again in order of keyness). A large number of these keywords can be summarised under four headings:

- 'female' items (e.g. *her*, *she*, *Miss*, *herself*, *sisters*),

- modal verbs (*could, must, might, would*),
- emphasisers/boosters (e.g. *very, such, really, certainly, perfectly*), and
- 'emotion' nouns (*feelings, engagement, acquaintance, attachment, attentions*).

From these categories it is possible to infer (still keeping in mind that this might be a pre-reading activity) that Austen's novels deal with the world of women, with issues related to obligation, permission, and probability, and with human emotions.

1. her	11. must	21. soon	31. have
2. she	12. been	22. much	32. always
3. be	13. every	23. nothing	33. would
4. very	14. such	24. really	34. exactly
5. Miss	15. feelings	25. am	35. attachment
6. Mrs	16. do	26. think	36. certainly
7. not	17. being	27. might	37. had
8. could	18. quite	28. engagement	38. attentions
9. herself	19. any	29. sisters	39. sister's
10. it	20. was	30. acquaintance	40. perfectly

Table 1: Keywords in JAC, excluding names of characters

Particularly striking here is the high frequency of different kinds of boosters (as opposed to hedges) and emphatic expressions. Since such expressions often mark humour or instances of ironic language use (cf. Barbe 1995: 22, Kreuz & Roberts 1995: 24, or Seto 1998: 241ff.), I will now briefly address this topic and present some results from a corpus-driven study on irony in JAC (cf. Oltmann 2004).

Austen and Irony – Collocations, Semantic Prosodies, and Deviations from the Norm

Instances of irony in texts are certainly not easy to spot, neither manually in a traditional reading approach, nor by means of any computer software, and especially not in older texts which lie outside our personal scope of experience:

Since irony is bound to intertextuality and social and stylistic conventions of language use, it is often difficult to detect instances and correctly interpret allusions and connotations, especially in historical periods (where readers may also be misled by anachronistic interpretations based on contemporary conditions). (Görlach 1999: 152)

Oltmann (2004: 45) also stresses this problem and notes that ironic instances, which come in many different shapes, "are hard to find using the tools of corpus linguistics because they follow few patterns, they are usually highly specific and creative, infrequent and context-based". Nevertheless, Oltmann in her corpus-driven study manages to establish several ironic devices and traces a number of items that often function as the source of irony or humour.

Central in her analysis are concordances, and the key analytic step consists in the comparison of the collocates and semantic prosodies of words in JAC on the one hand, and the collocates and semantic prosodies of the same words in the Novels Reference Corpus on the other hand. This comparison highlights how Austen's language differs from the language of a larger group of her contemporaries. In other words, Oltmann's analysis shows how Austen deviates from the norm of eighteenth/nineteenth century novel writing and how she creates meaning in her novels differently from contemporary authors.

The use of the word *recover* in JAC presents a nice example of this deviation from the norm, or as Mukařovský (1970: 42) calls it, "violation of the norm" (cf. also Barnbrook 1996, Louw 1993, and Louw 1997 on this issue). While *recover* is normally used in rather negative contexts with unfavourable semantic prosodies by Austen's contemporaries (see the concordance sample in

figure 12) – just as it is used in 'general' English today (people usually recover from something unpleasant),¹⁰ Austen uses collocations such as, "the agitating happiness of success", "the flutter of pleasure", or "a fit of good humour" in the context of *recover*, collocations that run counter the general negative pattern and thus create ironic or humorous effects.

N	Concordance
1	lict of penitence and passion. Ere he could recover from this agony of his spirits, the Pri
2	me such a shock,-I don't know when I shall recover from it. But I'm a sad, weak creatur
3	ou will be content to rest, and you will then recover from your delusion. You will perceiv
4	straw hat and blue plush gloves. He did not recover from the depression produced by th
5	ee. 'You must allow yourself a little time to recover from your first shock, my daughter,
6	ve related. My wife and my sister will never recover from their horror. I entreat you not t
7	; even Mrs. Mucklewrath, who had begun to recover from her hysterics, whimpered forth,
8	es that evening; but as she began a little to recover from her uneasiness at the disappo
9	n- choly despair, as my spirits could not yet recover from the violent shocks they had re
10	two yet; but he hardly thinks she will finally recover." "Has she mentioned me lately?"

Figure 12: The use of *recover* in NRC (a 10-line sample from the concordance)

N	Concordance
1	his friend. It was long before Fanny could recover from the agitating happiness of suc
2	understood him; and as soon as she could recover from the flutter of pleasure, excited
3	length closed her eyes. She could not yet recover from the surprise of what had happ
4	as they were gone, Elizabeth walked out to recover her spirits; or in other words, to dw
5	as been done, what has been attempted, to recover her?" "My father is gone to Londo
6	, I thank you," she replied, endeavouring to recover herself. "There is nothing the matte
7	house, to get sober and cool, and the other recover his temper and happiness when thi
8	a fit of good humour, from which he did not recover till after Eleanor had obtained his fo
9	over to any of them. She began at length to recover, to fidget about in her chair, get up,
10	h eagerness, that Emma might have time to recover-- "You may well be amazed. But it

Figure 13: The use of *recover* in JAC (a 10-line sample from the concordance)

Other unusual collocations that can be found in Austen's novels include "attack + watch" as in example (1), "attack + gallantry and compliment" as in (2), and "pleasure + mischief" as in (3).

¹⁰ Examples of such 'unpleasant' experiences in NRC are *agony*, *delusion*, *depression*, *shock*, and *horror*. A concordance based on the written part of the British National corpus highlighted collocates such as *injury*, *strain*, *wounds*, *depression*, and *anorexia*.

- (1) "Oh! Do not *attack* me with your watch. A watch is always too fast or too slow. I cannot be dictated to by a watch." (*Mansfield Park*)
- (2) For a day or two after the affront was given, Henry Crawford had endeavoured to do it away by the usual *attack* of gallantry and compliment, but he had not cared enough about it to persevere against a few repulses; (*Mansfield Park*)
- (3) Indeed she had no taste for a garden; and if she gathered flowers at all, it was chiefly for the *pleasure* of mischief – at least so it was conjectured from her always preferring those which she was forbidden to take. (*Northanger Abbey*)

Comparative cross-checks in JAC and NRC of randomly selected keywords from the JAC list also helped to identify a group of words that almost always express irony in Austen's novels, though usually not in the novels of her contemporaries. The semantically related words *fortune*, *glory*, *heroine*, *heroism*, *honour*, and *bustle* belong to this group (cf. (4) and (5) for illustrative examples from JAC).

- (4) This was all very promising; and, but for such an unfortunate fancy for having his hair cut, there was nothing to denote him unworthy of the distinguished *honour* which her imagination had given him; the *honour*, if not of being really in love with her, of being at least very near it, and saved only by her own indifference – (for still her resolution held of never marrying) – the *honour*, in short, of being marked out for her by all their joint acquaintance. (*Emma*)
- (5) But from fifteen to seventeen she was in training for a *heroine*; she read all such works as *heroines* must read to supply their memories with those quotations which are so serviceable and so soothing in the vicissitudes of their eventful lives. (*Northanger Abbey*)

Louw (1997: 244) observes that "the difference between the norm and features of the text is responsible for many of the 'devices' which give the reader so much pleasure", and indeed, instances of irony and humour in Jane Austen's novels certainly belong to these 'devices'. The corpus-driven identification of norm deviations in literary texts, however, can do more for readers and text analysts than indicate sources of reading pleasure or confirm intuitions about peculiar passages of a text. I would argue that the type of corpus-analytic exploration described in this essay can help us answer the central question how meaning is constructed in literary texts, and in what ways this construction of meaning differs from normal everyday language.

Surely, meanings that are created in literature are often rather subtle and not directly visible to the naked eye of the literary scholar or the student of literature (the proverbial 'reading between the lines' is not normally a straightforward activity), or, quoting Louw again,

[t]he belief is gradually becoming firmer that the question of *meaning can no longer be settled at first sight during the act of reading*, especially where first sight refers to a reading which is 'unassisted' by data. (Louw 2004: 1, my emphasis)

This is where corpora and concordance programs can help. As I hope the study of Austen's novels has shown, the skilful use of corpus-analytic techniques can contribute to making the invisible visible, and subtleties in literary language more noticeable.¹¹ What we need to do in the analysis and interpretation of literary texts is pay more attention to patterns and collocations in 'normal' natural language and compare them to the collocations and patterns in the more specialised literary language. That means that an investigation of the special or deviating patterns in a literary text should ideally be preceded by "a clear and comprehensive account of the established meaning-bearing patterns of the language" (Sinclair forthcoming). Larger reference corpora can help us identify such patterns with some precision.

¹¹ See also Stubbs' argumentation, this volume.

3.3 Corpus Analysis and Language Pedagogy

Having exemplified the use of corpora and corpus-analytic tools in linguistics and literary studies, I will now demonstrate how language pedagogy can also profit greatly from corpus work. I argue that students of English language didactics, most of them future teachers of English as a foreign language, should be acquainted with the basic steps in corpus analysis and that they should be familiar with some central direct and indirect applications of corpora in language teaching.

By 'direct application' I mean the use of corpora and concordances in the classroom. This approach is usually referred to as 'data-driven learning' (DDL). In a nutshell, the idea behind DDL is that learners are presented with corpus data, the exploration of which will lead them to discover regularities (and oddities) about the language they are learning. In the following I shall not, however, elaborate on this direct approach, which features learners as researchers, but provide an example of the indirect use of corpora in language teaching, which focuses on the pedagogical application of corpus research findings.

I made extensive use of the tools and methods of corpus analysis in a study of if-clauses in spoken British English and so-called 'school' English (cf. Römer 2004b). Starting from the observation that their use often presents a significant problem to learners, even on an advanced level, I examined whether if-clauses are treated differently in EFL teaching materials than they are actually used in natural spoken English. Of particular interest in this analysis was the sequence of tense forms, since this appears to cause particular difficulties for learners. I, therefore, determined the respective tense form sequence in altogether 211 if-clauses in a small and specialised corpus of spoken-type texts (dialogues, interviews, etc.) taken from two best-selling German EFL coursebook series and compared my findings to two if-clause datasets of similar size retrieved from the spoken part of the British National Corpus (BNC_spoken).

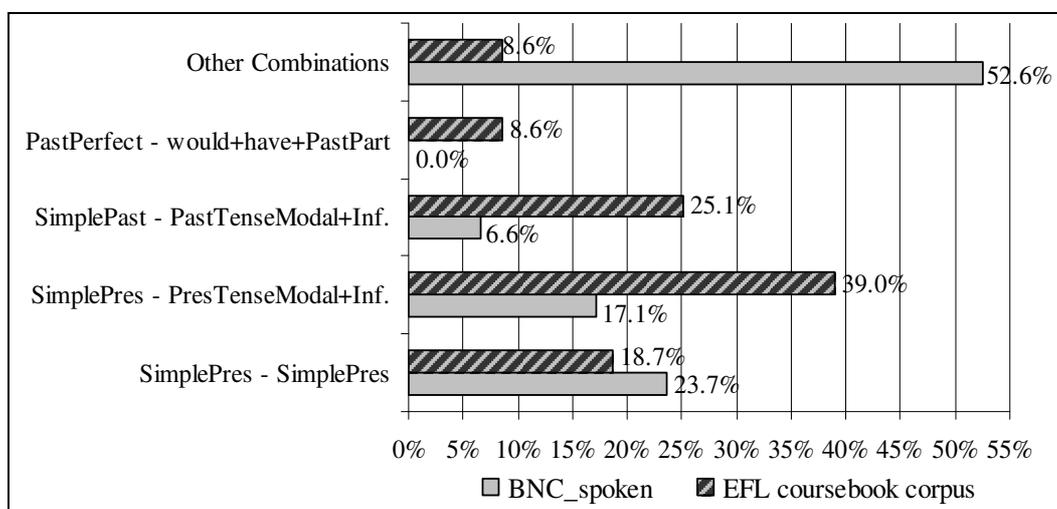


Figure 14: The distribution of if-clauses across tense form sequences in an EFL coursebook corpus and in BNC_spoken

Figure 14 illustrates the distribution of the four most common tense form combinations in the coursebook if-clauses and compares the determined shares to the respective results based on BNC_spoken data. We note some significant distributional differences between the two types of corpora. There is a clear overuse in EFL coursebook if-clauses of the three combinations that are usually introduced in the following way (on the basis of invented examples) and typically referred to as if-clauses 'type 1', 'type 2', and 'type 3' in the teaching materials:

- simple present (if part) – present tense modal (usually *will*) + infinitive (main part), as in "If I win the lottery, I will buy a house in Tuscany."
- simple past (if part) – past tense modal (usually *would*) + infinitive (main part), as in "If I won the lottery, I would buy a house in Tuscany."
- past perfect (if part) – past tense modal (*would*) + *have* + past participle (main part), as in "If I had won the lottery, I would have bought a house in Tuscany."

On the other hand, we observe a significant underuse of the 'simple present – simple present' sequence and of a number of other combinations that frequently occur in if-clauses in natural speech.

On the basis of these and other attested differences between real spoken English and 'school' English, I made a number of suggestions in what ways teaching materials could be changed so that they mirror more closely how language is really used, and not how writers' materials imagine it to be.¹² This study shows that corpus analysis can contribute to an improvement of teaching materials and that corpora can play a vital role in language pedagogy.

4 Conclusion: The Importance of Corpus Literacy in English Studies

In the introduction to this chapter I formulated the questions, "Can corpus analysis help to improve interdisciplinarity in English Studies?" and "Can corpora contribute to a reconciliation of literature, linguistics, and pedagogy?" By means of applying some central corpus-analytic steps in these different branches of our discipline, I hope to have shown that the answer to both questions has to be a definite "Yes."

I regard corpus-analytic techniques as multi-purpose strategies with an immense potential to enhance all sorts of textual analysis and to confirm or contradict our intuitions about patterns and meanings in literary and non-literary language. Since we all work with texts in some way or other, no matter whether we teach *if*-clauses, interpret Jane Austen's novels, or investigate modal verbs, it appears only natural to share some of the strategies we use, thus assigning universal status to them. Figure 15 aims to illustrate the universality of corpus-analytic text exploration in the different intersecting sub-branches of English Studies. I see corpus analysis as a key skill that ought to reach all parts of the discipline as it can be equally applied in linguistics, literary studies, and language pedagogy.

¹² A similar comparative analysis, though on a considerably larger set of data, was also carried out on English progressive forms (cf. Römer 2005 and forthcoming).

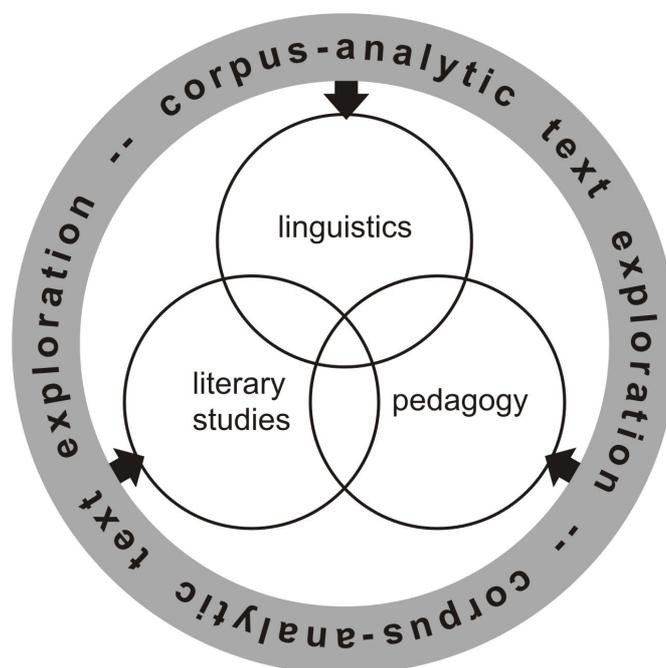


Figure 15: Corpus analysis in English Studies

Getting back to the perspective of our students, I think we have to consider what they need in order to succeed in their studies. What should they be equipped with so that they will be able to handle problems from different parts of the subject? My suggestion would be that we equip our students with a tool box, containing skills that are transferable from problem to problem across sub-disciplines. One of these skills should relate to the knowledge of making use of corpora and software for corpus analysis. I would like to call this particular skill 'corpus literacy' and regard it as a key literacy in English Studies.¹³ The acquisition of a corpus literacy including a certain amount of 'hands-on' experience in computer-aided text analysis may open new perspectives for students. It can make them more independent and show them that, even at an early stage in their studies, they can be researchers themselves. Also, as convincingly discussed by

¹³ In the context of English language teaching, Mukherjee (2002: 153) also mentions "corpus literacy" as an important skill that pupils have to acquire if they are to profit from the use of corpora and corpus-analytic software in the foreign language teaching classroom.

Kettemann and Marko (2004), the use of corpora and corpus-analytic techniques in English Studies helps to strengthen students' language awareness, in particular their discourse awareness, i.e. it enables them to distinguish more clearly between different types of texts or discourses within the discipline.

N	Concordance
1	of the transaction. I find great difficulty in reconciling that proposition with the proposi
2	erged. For months they had had difficulty in reconciling the accounts until they realized t
3	er person and which she/he has difficulty in reconciling with her/his own self-image. Th
4	cause problems because of the difficulty in reconciling their continuing duty to their ex-
5	and justice. Mansfield saw the difficulty in reconciling the two principles, but thought t
6	in such a context, the greater the strain in reconciling internationalism and nationalism.
7	e;lite, but in total Japan did not succeed in reconciling the most articulate in Korea to th
8	tal failing symptomatic of the impossibility of reconciling communist ideology with the arti
9	ant taking on board the awkward problem of reconciling, within a reconstructed union fra
10	y a Christian. There arises the problem of reconciling a religion which has a unique C
11	lphia research acknowledges the problem of reconciling within a unitary study the need f
12	to be wrong, we are left with the problem of reconciling new observation and established
13	w lights in each main gable, the problem of reconciling this glazing with the added upp
14	n the 1920s drew attention to the problem of reconciling this emerging statehood with the
15	quo. This concern highlights the problem of reconciling professional practice and standa
16	are involved. A further possible way of reconciling the two accounts is to say that
17	it; and whether there are alternative ways of reconciling security and control with human
18	time throughout his later life to find ways of reconciling quantum mechanics with a more
19	Much of his fiction is spent dramatising or reconciling the dichotomy. Wells employed

Figure 16: The use of *reconciling* in written English (sample from a BNC_written concordance)

If we look at the concordance sample in figure 16, we note that in general English language use the word *reconciling* carries a rather negative semantic prosody and that "difficulty in reconciling" and "problem of reconciling" count among the main collocations. Despite the unfavourable context in which *reconciling* apparently tends to occur, I believe that there are possible ways of reconciling English Studies, and, in fact, we find evidence for this in lines 16 to 18 of the displayed concordance. Like Mair (2002: 126) I am convinced that corpora, defined in a wide sense as computerised text collections, and corpus tools can be beneficially used in all parts of our discipline which would then "pave the

way for philology, the scholarly analysis of texts, to re-establish itself in the digital age". If this can be achieved, a reconciliation of the different facets of English Studies cannot be far.

References

- Barbe, Katharina. *Irony in Context*. Amsterdam: Benjamins, 1995.
- Barlow, Michael. *MonoConc Pro*. Houston: Athelstan, 2000.
- . "Software for Corpus Access and Analysis" in *How to Use Corpora in Language Teaching*, ed. John McH. Sinclair. Amsterdam: Benjamins, 2004. pp. 205-21.
- Barnbrook, Geoff. *Language and Computers*. Edinburgh: Edinburgh University Press, 1996.
- Biber, Douglas, Susan Conrad & Randi Reppen. *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press, 1998.
- Bublitz, Wolfram. "Semantic Prosody and Cohesive Company: 'Somewhat Predictable'," *Leuvense Bijdragen (Leuven Contributions in Linguistics and Philology)* 85/1-2 (1996), 1-32.
- Esser, Jürgen. "Collocation, Colligation, Semantic Preference and Semantic Prosody: New Developments in the Study of Syntagmatic Word Relations" in *Words, Lexemes, Concepts – Approaches to the Lexicon: Studies in Honour of Leonhard Lipka*, ed. Wolfgang Falkner & Hans-Jörg Schmid. Tübingen: Narr, 1999. pp. 155-65.
- Görlach, Manfred. *English in Nineteenth-Century England. An Introduction*. Cambridge: Cambridge University Press, 1999.
- Hunston, Susan. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press, 2002.
- Hunston, Susan & Gill Francis. *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: Benjamins, 2000.

- Kennedy, Graeme. *An Introduction to Corpus Linguistics*. London: Longman, 1998.
- Kettemann, Bernhard & Georg Marko. "Can the L in TALC Stand for Literature?" in *Corpora and Language Learners*, ed. Guy Aston et al. Amsterdam: Benjamins, 2004. pp. 169-93.
- Klages, Monika & Ute Römer. "Translating Modal Meanings in the EFL Classroom" in *Language: Context and Cognition: Papers in Honour of Wolf-Dietrich Bald's 60th Birthday*, ed. Sybil Scholz et al. Munich: Langenscheidt-Longman, 2002. pp. 201-16.
- Kreuz, Roger J. & Richard M. Roberts. "Two Cues for Verbal Irony: Hyperbole and the Ironic Tone of Voice," *Metaphor and Symbolic Activity* 10/1 (1995), 21-31.
- Krishnamurthy, Ramesh. "The Macrocosm and the Microcosm: The Corpus and The Text" in *Linguistic Approaches to Literature: Papers in Literary Stylistics* [Discourse Analysis Monograph 17], ed. Jonathan Payne. Birmingham: The University of Birmingham, 1995. pp. 1-16.
- Louw, Bill. "Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies" in *Text and Technology. In Honour of John Sinclair*, ed. Mona Baker et al. Amsterdam: Benjamins, 1993. pp. 152-76.
- . "The Role of Corpora in Critical Literary Appreciation" in *Teaching and Language Corpora*, ed. Anne Wichmann et al. London: Longman, 1997. pp. 240-51.
- . "Truth, Literary Worlds and Devices as Collocation." Presentation abstract of a keynote lecture held at the 6th Teaching and Language Corpora (TaLC 6) conference, 6-9 July 2004, Granada, Spain.
- Mair, Christian. "Empowering Non-native Speakers: The Hidden Surplus Value of Corpora in Continental English Departments" in *Teaching and Learning by Doing Corpus Analysis*, ed. Bernhard Kettemann & Georg Marko. Amsterdam: Rodopi, 2002. pp. 119-30.
- McEnery, Tony & Andrew Wilson. *Corpus Linguistics*. 2nd ed. Edinburgh: Edinburgh University Press, 2001.
- Meyer, Charles F. *English Corpus Linguistic: An Introduction*. Cambridge: Cambridge University Press, 2002.

- Mukařovský, Jan. "Standard Language and Poetic Language" in *Linguistics and Literary Style*, ed. Donald C. Freeman. New York: Holt, Rinehart and Winston, 1970. pp. 40-56.
- Mukherjee, Joybrato. *Korpuslinguistik und Englischunterricht. Eine Einführung*. Frankfurt: Lang, 2002.
- Oltmann, Katrin. "The Pleasure of Mischief: A Corpus-Driven Approach to Irony and Humour in the Novels of Jane Austen." Unpublished research paper, University of Cologne, Germany, 2004.
- Partington, Alan. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: Benjamins, 1998.
- . "'Utterly Content in Each Other's Company': Semantic Prosody and Semantic Preference," *IJCL* 9/1 (2004), 131-56.
- Reppen, Randi. "Review of MonoConc Pro and WordSmith Tools," *Language Learning & Technology* 5/3 (2001), 32-6.
- Römer, Ute. "A Corpus-driven Approach to Modal Auxiliaries and their Didactics" in *How to Use Corpora in Language Teaching*, ed. John McH. Sinclair. Amsterdam: Benjamins, 2004a. pp. 185-99.
- . "Comparing Real and Ideal Language Learner Input: The Use of an EFL Textbook Corpus in Corpus Linguistics and Language Teaching" in *Corpora and Language Learners*, ed. Guy Aston et al. Amsterdam: Benjamins, 2004b. pp. 151-68.
- . *Progressives, Patterns, Pedagogy. A Corpus-driven Approach to Progressive Forms, Functions, Contexts and Didactics*. Amsterdam: Benjamins, 2005.
- . "Looking at *looking*: Functions and Contexts of Progressives in Spoken English and 'School' English" in *The Changing Face of Corpus Linguistics*, ed. Antoinette Renouf & Andrew Kehoe. Amsterdam: Rodopi, forthcoming.
- Scott, Michael. *WordSmith Tools* (Version 3.0). Oxford: Oxford University Press, 1999.
- Seto, Ken-ichi. "On Non-echoic Irony" in *Relevance Theory: Applications and Implications*, ed. Robyn Carston & Seiji Uchida. Amsterdam: Benjamins, 1998. pp. 239-55.