What is corpus linguistics? What the data says

Article (Published Version)

What is *corpus linguistics*? What the data says

Charlotte Taylor
University of Siena

1 Introduction: Explicit definitions

Stubbs (2006), in his state of the art overview, draws attention to the frequent reticence or vagueness of corpus analysts in discussing their operational methods within a scientific context, (a context addressed in detail in Partington forthcoming). This lack of clarity in discussing the methodological framework employed is, perhaps, most surprising given the way in which corpus linguistics situates itself within a scientific frame, and lays such claims to a scientific nature.

This brief paper, then, addresses the question posed in its title, namely, "What is corpus linguistics?" – is it a discipline, a methodology, a paradigm or none or all of these? – but does not attempt to offer any definitive answers. Rather, the aim is to present the reader with a number of observations on how corpus linguistics has been construed in its own literature and then to leave the question open, in the hope of stimulating further discussion. The study takes the specific term *corpus linguistics* and looks at how it is defined and described both explicitly and implicitly in a variety of relevant sources.

There is no shortage of overt discussion among theorists of what corpus linguistics is or should be. However, at the same time, to the casual observer or new arrival there might also appear to be a bewildering variety of definitions and descriptions. Aarts, one of the founding fathers, seems to have anticipated this. Aarts and Meijs (1984), as the first book dedicated to the subject, is often identified as the source of the term c*orpus linguistics*, although the term had in fact been used previously, for example, in Aarts and van den Heuvel (1982). On the Corpora List, Aarts is reported as commenting that the term was coined with some hesitation "because we thought (and I still think) that it was not a very good name: it is an odd discipline that is called by the name of its major research tool and data source. Perhaps the term has outlived its usefulness by now".[1] This raises one of the recurrent concerns over talking about *corpus linguistics*, and may account for the preference for alternatives.

In terms of what corpus linguistics 'is', not only have various definitions been offered, but alternatives have been explicitly addressed and rejected. These include, as we shall see: *corpus linguistics* is a *tool*, a *method*, a *methodology*, a *methodological approach*, a *discipline*, a *theory*, a *theoretical approach*, a *paradigm* (theoretical or methodological), or a combination of these.

In 1992, Leech argued that "computer corpus linguistics defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject" and goes on to describe the characteristics of computer corpus linguistics as a new paradigm (Leech 1992: 106). Similarly, Stubbs 1993, rejects the limited definition of corpus linguistics as a methodology, and, commenting on Sinclair 1991, he notes that "[i]n this vision of the subject, a corpus is not merely a tool of linguistic analysis but an important concept in linguistic theory" (1993: 23–24). Teubert (2005) also emphasises the theoretical conceptualisation and describes corpus linguistics as "a theoretical approach to the study of language" (2005: 2).

The notion of corpus linguistics as a paradigm is taken up by Gries, but the methodological conceptualisation is favoured, as he states that "[o]ver the past few decades, corpus linguistics has become a major methodological paradigm in applied and theoretical linguistics." (2006a: 191). In 2001, Tognini-Bonelli described corpus linguistics as a "pre-application methodology" which possesses "theoretical status" (2001: 1). Similarly, Mahlberg describes corpus linguistics as "an approach to the description of English with its own theoretical framework" (2005: 2), and to emphasise this employs the term "corpus theoretical approach" (2005, 2006). In directly addressing the issue she sees the difference of perception as stemming from the type of corpus linguistics which the researcher practices: "[t]here is still disagreement on whether corpus linguistics is mainly a methodology or needs its own theoretical framework. Advocates of corpus-driven approaches to the description of English claim that new descriptive tools are needed to account for the situation of real text, and ideas of theoretical frameworks to accommodate such tools have started to emerge." (2006: 370). Thompson and Hunston (2006) state that "[a]t its most basic corpus linguistics is a methodology that can be aligned to any theoretical approach to language" (2006: 8). However, they go on to describe two major theories which have come out of corpus linguistics. First of all, that meaning is not located in single words, but in 'units of meaning' in Sinclair's terminology, and consequently that communicative discourse unfolds largely as a series of semi-fixed phrases (2006: 11–12).

McEnery, Xiao and Tono, note that as "corpus linguistics is a whole system of methods and principles of how to apply corpora in language studies and

teaching/learning, it certainly has a theoretical status. Yet theoretical status is not theory in itself" (7–8: 2006); they therefore conclude that corpus linguistics is a methodology. Corpus linguistics is also defined as a methodology in McEnery and Wilson (1996) and Meyer (2002), and as "an approach or a methodology for studying language use" in Bowker and Pearson (2002: 9). However, McEnery and Gabrielotos note that "[c]orpus linguistics may be viewed as a methodology, but the methodological practices adopted by corpus linguists are not uniform" (2006: 44), and they illustrate how such methodological differences are driven by theoretical considerations. Teubert (2005) also comments on the diversity of methods, and states that "[c]orpus linguistics is not in itself a method: many different methods are used in processing and analysing corpus data. It is rather an insistence on working only with real language data taken from the discourse in a principled way and compiled into a corpus" (2005: 4).

Aarts (2002), Teubert (2005), and Williams (2006), among others, describe corpus linguistics as a discipline. This, in turn, raises the further question of precisely what type of discipline, for example, while Stubbs (1993: 3) describes linguistics as an "applied social science", Teubert states that "[l]inguistics is not a science like the natural sciences whose remit is the search for 'truth'. It belongs to the humanities, and as such it is a part of the endeavour to make sense of the human condition." (2005: 7).

In defining and characterising *corpus linguistics*, others have emphasized the hard-science credentials of corpus linguistics. For example, McCarthy describes corpus linguistics as representing "cutting edge change in terms of scientific techniques and methods" (2001: 125), and Stubbs explicitly parallels corpus linguistics and science, noting that "[g]eologists are interested in processes which are not directly observable because they take place over vast periods of time […] Corpus linguists are interested in processes which are not directly observable because they are instantiated across the language use of many different speakers and writers" (2001: 243). Another indication of the importance attached to the scientific method in corpus linguistics may be gleaned from a comparison with the applied sciences section of the BNC: WordSmith KeyWords for the corpora about corpus linguistics used in the present paper (see section 2 below) included: *repetition, empirical, statistical, methodology, data, quantitative* and *qualitative.*

Interestingly, it is on the claim to scientific method that Chomsky criticised corpus linguistics, stating "[m]y judgment, if you like, is that we learn more about language by following the standard method of the sciences. The standard method of the sciences is not to accumulate huge masses of unanalyzed data and to try to draw some generalization from them" (2004: 97). Chomskyan linguis-

tics, in turn, is frequently commented on, and criticised in corpus linguistics. Carter claims that it displays "no interest in language beyond the level of the sentence, there is no recognition that authentic data is of any significance and there is no acceptance that studies of large corpora or real language in use play any part in descriptive theories of language. Most significantly, too, there is a clear sense that the analysis of meaning is not a primary purpose." (2004: 2). Sinclair also criticised introspective linguistics by referring to science, noting that "one does not study all of botany by making artificial flowers" (1991: 6). More recently, Teubert and Krishnamurthy (2007) describe corpus linguistics as *parole*-linguistics opposing it to the *langue*-linguistics of Saussure and Chomsky among others. This mode of definition by reference to alterity appears to be a common feature of corpus linguistics, and even within the two small corpora of research papers used in this study, four articles refer to Chomsky.

It has also been stressed that the variety of stance in describing *corpus linguistics* is no bad thing. Differing interpretations are to be expected, not only because, as Hoey (1993: vi) notes, the scientific status of linguistics itself has been much discussed over the years, but also because corpus linguistics is evolving: research is being carried out in a wide variety of ways and covers a range of topics. Furthermore, the nature of corpus linguistics means that the corpus linguist may be corpus designer, compiler or analyst – and indeed is often all three – and each of these might have a different vision of what the enterprise entails. Such differing interpretations are to be welcomed: as Teubert states in presenting his version of corpus linguistics "[o]nly if the discourse of corpus linguistics remains controversial and pluralist will there be progress" (2005: 13). However, for such progress to take place, there needs to be discussion of the different interpretations, on how they overlap, contrast and interact. Moreover, an appreciation of the positive variety of views at the level of the academic community should not be confused with an acceptance of vagueness on the part of the individual researcher.

The initial aim of this study was simply to collect definitions of corpus linguistics. What came out of this preliminary exploration was that while there is a multiplicity of views available from some of the most influential corpus theorists, there is little explicit discussion of what they are involved in by the majority of practitioners of corpus linguistics, and few works contain clarifications of the concepts with which the researcher is working. Secondly, the term *corpus linguistics* itself seems often to strike practitioners as problematic, and even a quick look at the titles of some of the most influential textbooks on corpus linguistics reveals that the term is relatively under-used. Definitions frequently

refer not to *corpus linguistics*, but to *corpus/corpus-based/corpus-driven/corpus assisted + analysis/approach/study* etc.

Tognini-Bonelli (2001), in particular, argues for recognition of Corpus-driven Linguistics (CDL) as a discipline within corpus linguistics, a position supported in Römer (2005). Others have adopted the term *corpus-assisted* research and emphasize how corpus-analysis tools and techniques can be integrated and enhanced by other approaches, other 'ways into' to the data-set under examination. These include sampling, by reading or watching texts from the corpus, a process which can help provide a feel for how things are done linguistically in the discourse-type, comparison and 'para-replication' of findings from one corpus with others, and using other non-corpus means of acquiring information about participants and practices in the discourse type (Partington 2004, forthcoming; Morley and Bayley forthcoming). Following Fillmore (1992), they also stress the need to exploit the productive interplay of intuition, data-observation and introspection.

The aims of this paper are, therefore, simply to look at the ways in which the specific term *corpus linguistics* is used and defined in practice, and, in so doing to offer up some food for thought to the reader.

2 The data

2.1 The Corpus Linguistics Research Articles (CLRA) corpus

In order to look at the perceptions of researchers working with/within corpus linguistics, a small corpus of research articles, the Corpus Linguistics Research Articles (CLRA) corpus, was compiled from relevant online journals. Rather than referring exclusively to book-length publications, research articles were selected as it was felt that they could represent a wider range of researchers. Articles were downloaded from various online publishers if they included the term *corpus linguistics* in the keywords (37 articles), or, in the case of articles which did not include keywords in their format, if the term *corpus linguistics* appeared in the title or abstract (10 articles). The reason for this selection was to ensure that the corpus contained articles which explicitly 'declared' that they were about *corpus linguistics*.[2] The journals were not pre-selected in order to try and avoid defining corpus linguistics prior to looking at all the available data.

In total, 47 articles were collected from 20 different journals (see Appendix), over the time period 1996–2007, forming a corpus of 463,489 tokens. The size of the corpus was simply determined by the journals to which the university (Siena) had a subscription, and whether the journals allowed for conversion to a .txt format, which was necessary for subsequent analysis using *WordSmith Tools*

3.0 (Scott 1999). The texts were marked up to allow for identification of title, keywords, abstract, references and main body. Main body was defined as from the start of the introduction to the end of the conclusion, excluding footnotes, and within the main body, the sections introduction and conclusion were also identified.[3]

2.2 Corpora for comparison

To allow for comparison and extension, two additional corpora were compiled. The first consists of research articles about conversation analysis, which was chosen as, like corpus linguistics, it clearly refers to a community of practice within linguistics. The Conversation Analysis Corpus contains 58 research articles (630,049 tokens) which were selected from online journals in the same way as those for the CLRA corpus: they were downloaded if they contained the term *conversation analysis* in the keywords and could be converted to .txt.

The second comparison corpus, CONF, consists of papers presented at the conference *Corpus Linguistics 2005* (held in Birmingham July 14–17, 2005) and contains 69 articles (349,942 tokens). All the articles which could be converted to .txt were downloaded from the *Proceedings for the Corpus Linguistics Series 2005* website.[4] While keywords were not part of the article format, it was felt that, as they had been presented at the conference *Corpus Linguistics 2005*, they were inherently declaring that they were about corpus linguistics.

All the corpora were analysed using *WordSmith Tools 3.0*.

3 Analysis

3.1 The Corpus Linguistics Research Articles (CLRA) corpus

The frequency and distribution of the term *corpus linguistics* in the CLRA corpus proved to be quite intriguing. As can be seen from the dispersion plot (Figure 1), the occurrences of the term *corpus linguistics* are unequally distributed both among and within the research articles. Ten articles contain only one occurrence of *corpus linguistics* which, given the way in which the corpus was compiled, was of course in the title, keywords, or abstract. Surprisingly, of the 47 research articles approximately half – 23 – did not refer to *corpus linguistics* in the main body of the text at all. Given that 20 of these 23 included the term *corpus linguistics* in the keywords, this may raise questions either about the function of keywords, or about the use of the term *corpus linguistics*.

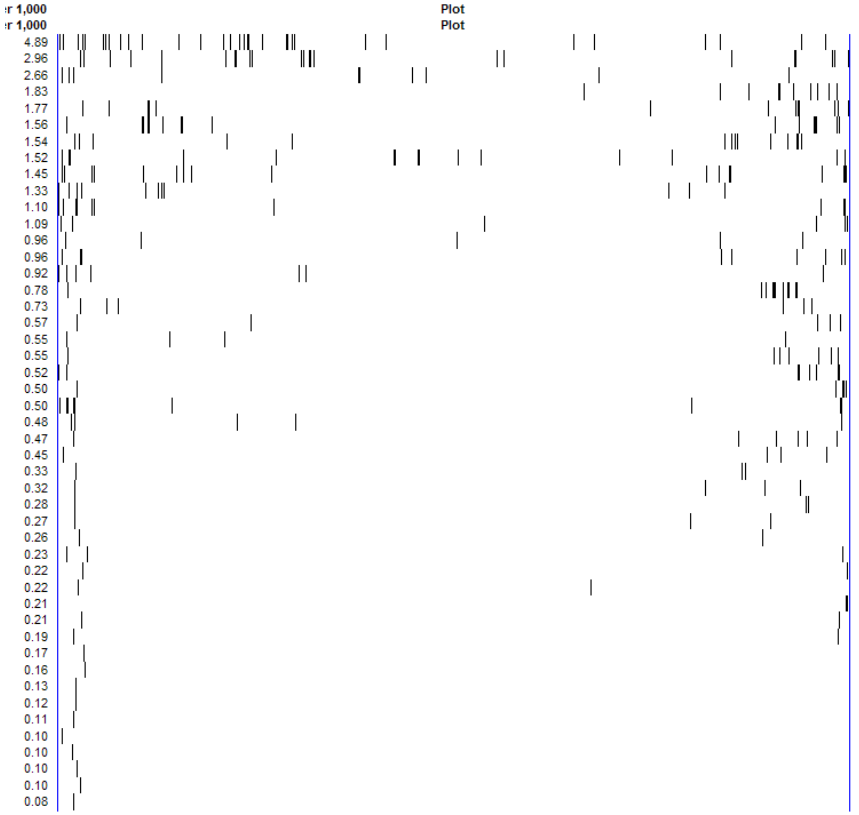

Figure 1: Dispersion plot for corpus linguistics *in the CLRA corpus*

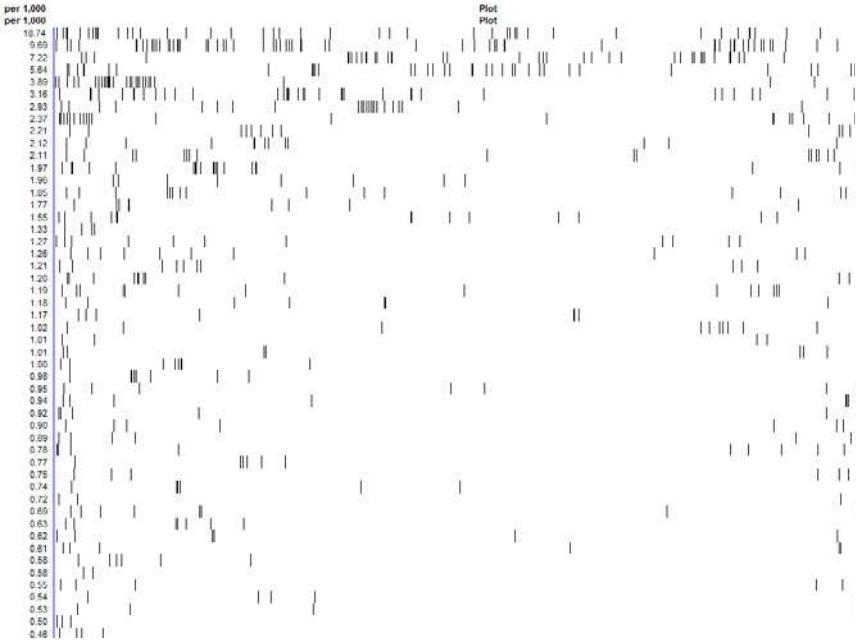

Figure 2: Dispersion plot for conversation analysis *and* CA *in the Conversation Analysis corpus*

For ease of comparison, in Figures 1 and 2 the dispersion plots have been placed in sequence, and the differences in frequency of the search terms can be clearly seen. In the Conversation Analysis Corpus, *conversation analysis* occurred 388 times (0.6 ptw)[5] and its abbreviation *CA* occurred 434 times (0.7 ptw). This compares to an occurrence of 0.7 ptw of the term *corpus linguistics* in CLRA. The dispersion plot reflected similar tendencies within the articles from both corpora, as occurrences clustered towards the beginning and the end. However, as can be seen, in the Conversation Analysis corpus there were no articles with only one occurrence, in contrast with the CLRA corpus.

Figure 3 shows the distribution of *corpus linguistics* within the research articles in more detail. Occurrences in the title, keywords and abstract accounted for 19 per cent of the total, and 38 per cent of occurrences are found in the bibliographies,[6] with just 43 per cent in the main body of the article. Furthermore, 9 of

the occurrences within the main body were in subheadings, which reflects the overall tendency in the corpus to mark the text as being 'about' *corpus linguistics*.

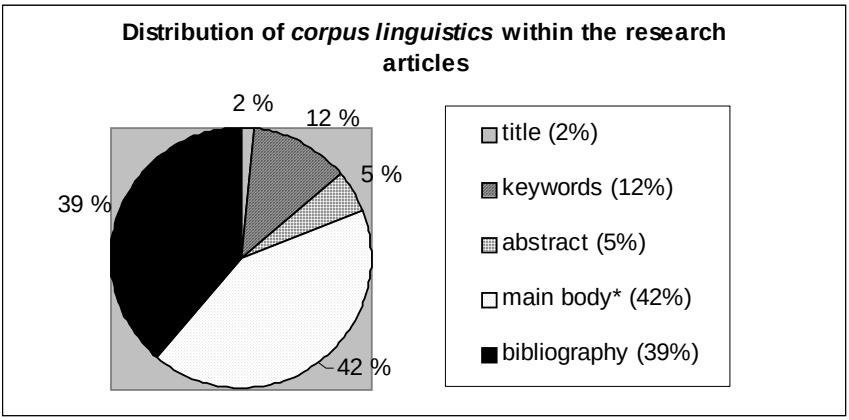

Figure 3: Distribution of corpus linguistic*s within the research articles (CLRA)*

Very similar results were found in the CONF corpus. The term *corpus linguistics* was only present in 37 of the 70 papers and, out of the 107 occurrences, bibliographic references made up approximately 60 per cent. Excluding these bibliographic references only 16 of the 70 papers actually included the term *corpus linguistics*.

Returning to the concentration observed towards the beginning and the end of the research articles, as seen in Figures 1 and 2, a closer analysis showed that there were 34 occurrences of our search term (in 16 articles) in the introductions and 12 (in 9 articles) in the conclusions, 27 per cent and 10 per cent respectively of the total. These concentrations were higher in the CONF corpus: within the main body there were 38 occurrences, of which 16 (42 per cent) were found in the introductions and 7 (18 per cent) in the conclusions.

3.2 Right collocates of corpus

It could be argued that the term *corpus linguistics* occurs less frequently than expected because there are so many alternatives available, as the R1 collocates of *corpus* from the CLRA corpus illustrate (see Table 1). For instance the collocates *based* and *driven* may be seen as adding an additional layer of delicacy by specifying a type of corpus linguistics. However, the terms rarely occur in the

form *corpus + based/driven + linguistics*; the most frequent clusters in this corpus were *corpus-based study/studies* and *corpus-driven research.* This may indicate unease with the term *corpus linguistics*, or that other terms are employed since *corpus linguistics,* as used in the keywords, indicates the field or discipline and academic orientation of the research rather than the research activity which is also described, for example, in the collocates of *corpus analysis*, and *corpus research*:

Table 1: Selected R1 collocates of corpus from the main body of articles in CLRA[7]

	TOTAL	R1	No. of articles
BASED	196	154	27
LINGUISTICS	138	124	24
DATA	136	78	16
ANALYSIS	74	28	14
DRIVEN	37	24	5
LINGUISTIC	37	23	10
EVIDENCE	33	21	11
APPROACH	50	17	5
STUDIES	59	17	10
RESEARCH	77	16	13

3.3 Comparing the research articles that contain corpus linguistics *in the main body of the articles (CLRA-Y) with those that do not (CLRA-N)*

Although the corpus used here is clearly too small for any firm generalisations to be made, the results of the WordSmith KeyWords comparison indicate that there seem to be two slightly different discourse communities represented in this corpus.[8] For example, Key 3-word clusters for the CLRA-Y sub-corpus include *British National Corpus* (Keyness 31.4) and *Bank of English* (28.4). It is of particular interest to note that Key 2-word clusters for the CLRA-Y sub-corpus include *corpus-based* (39.4), and *corpus-driven* (35.8) which excludes the possibility that *corpus linguistics* is simply substituted by these terms in the articles which do not use *corpus linguistics* in the main body. Other Key 2-word clusters

for CLRA-Y include *corpus of* (35.8), and *concordance lines* (46.7), as well as *semantic prosody* (132.3), *the text* (56.8), *text analysis* (46.3), *of meaning* (31.3), and *field of* (24.5). Keywords for CLRA-Y referring to people included: *Sinclair*, *Stubbs*, *Louw*, and *Hunston*,[9] the only author in the Keyword list for CLRA-N was *Lakoff*.

Other keywords, hinting at differences between the two discourse communities, are shown in Table 2:

Table 2: Selected Key words of CLRA-Y compared to CLRA-N

CLRA-Y		**CLRA-N**	
ANALYSES	SEARCHES	CITATION	MODAL
ANALYSIS	SEMANTIC	CONCEPTUAL	PROCEDURAL
ASSOCIATIONS	SEMANTICS	DOMAIN	RELATED
COLLOCATES	SET	EXPRESS	RELEVANCE
COMPUTER	SOCIAL	EXPRESSION	REPETITION
CORPUS	SOFTWARE	EXPRESSIONS	STANDARD
EMPIRICAL	STRUCTURES	EXTRACTED	STATISTICAL
LANGUAGE	TECHNOLOGY	FORMULAIC	TOKENS
LINGUISTICS	TEXT	LEXICON	TRANSCRIPTION
MEANING	TEXTS	LIKELIHOOD	UTTERANCE
MILLION	WORD	MAPPING	

It may also be of interest to note that, in both CLRA and CONF, articles which did not use the term *corpus linguistics* were also less likely to employ the premodifier *corpus linguistic* or refer to *corpus linguists*, as illustrated in Table 3:

Table 3: Occurrences of *corpus linguistic* and *corpus linguist/s* in the sub-corpora of CLRA and CONF

	CLRA-Y		**CLRA-N**		**CONF-Y**		**CONF-N**	
	freq	ptw	freq	ptw	freq	ptw	freq	Ptw
corpus linguistic	16	0.088	3	0.015	3	0.014	1	0.005
corpus linguist/s	12	0.066	2	0.010	15	0.072	1	0.005

3.4 Widening the database: Interrogating WebCorp

In order to try and widen the research database, WebCorp[10] was used to collect examples from the web of the sequence corpus+linguistics+is. The concordance lines below show the examples which were followed by an indefinite article. As can be seen, the variety of interpretations described above is reflected in the results, which have been ordered thematically:

1.- Introduction	*Corpus Linguistics is* a **branch of Linguistics** which
edical translators	*Corpus Linguistics is* a **branch of linguistics** that
UCTION	*corpus linguistics is* a **branch of linguistics** in which
research. Firstly	*Corpus Linguistics is* a growing **discipline** that
's 60th birthday	*corpus linguistics is* a relatively new **subject empirical linguistics**,
of which	*Corpus linguistics is* a **representative,** sets itself
n general I think	*corpus linguistics* is an interesting **field** for us to look
rpus linguistics	*Corpus Linguistics is* **an applied field of linguistics**,
the time. By the way	*Corpus Linguistics is* **a field of linguistics** where
niversity of Birmingham	*Corpus linguistics* is a relatively new **field**, pioneered
s (of which more later…	*Corpus linguistics is* a growth **area**, and there are structure
or language use	*corpus linguistics is* a broader **concept** that can be
age Change *Quantitative*	*Corpus linguistics is* a newly emerging **empirical fra**
possible. Why? Because	*Corpus Linguistics is* **an empirical science**, in which
umber of linguistic fields	*corpus linguistics is* **a science** and **approach** which *
he language of literature	*corpus linguistics is* also **an empirical approach** to
Einführung . Peter Lang	*corpus linguistics is* **an empirical approach** to
in particular, with SFL	*Corpus linguistics is* **an empirical approach** to the
guage texts. *Multilingual*	*Corpus Linguistics is* a new **approach**, based on the
uthor) "Simply speaking	*Corpus linguistics is* **an approach or a methodology** *
structure or language use	*Corpus Linguistics is* a broader **concept**, a **methodology**
INTRODUCTION	*Corpus linguistics is* data-driven **methodology** for
uthors. We maintain that	*Corpus linguistics is* a **methodology**. While the fact
paper takes the view that	*corpus linguistics is* a **methodology** which can by
at is corpus linguistics (I	*corpus linguistics* is a **method** of carrying out e as a social
phenomenon	*corpus linguistics" is* **a practice, rather than a theor**
rpus analysis, and to me	*corpus linguistics* is an extremely valuable **tool**. I
eaking, this implies that	*corpus linguistics is* an important **tool** for work within

(WebCorp)[11]

The definitions range from *branch of linguistics* and *field* to *tool*. From the concordance lines it seems that the 'non-scientific' term *approach* is one of the most nebulous definitions, as it appears together with both descriptions of *corpus linguistics* as a *science* in itself and together with definitions of corpus linguistics as a *methodology* (concordance lines marked with an asterisk).

It is also interesting to note the frequency of *empirical* in the descriptions as this was also one of the Keywords for the CLRA-Y sub-corpus compared to CLRA-N (Table 2), as well as a Keyword for the CLRA corpus in the comparison with the Applied Science Writing section of the BNC, suggesting that this is a key characteristic in constructing the identity of corpus linguistics.

3.5 Defining and describing corpus linguistics *in the CLRA Corpus*

The search for definitions was then repeated on the CLRA corpus, which contains only published articles and could therefore be considered a more careful reflection of the perceptions of researchers working with/in *corpus linguistics*. The concordance lines below show the results for the same sequence of *corpus + linguistics + is* as retrieved from *WebCorp*. As can be seen, there were few explicit definitions, although the variety seen above appears to be present in the CLRA corpus too:

Corpus linguistics is now an established **field** with a growing body of
Corpus linguistics is a relatively new **field** of inquiry.
Corpus linguistics is a **methodology** which can be described as a study of
corpus linguistics is **not** **"a domain of study"**, but rather ''**a methodological**
corpus linguistics is **a means** of studying and describing language use which
corpus linguistics is **the study of** language through corpus-based or corpus-

(CLRA, 4 files)

The pattern in the last line, describing *corpus linguistics* as *the study of*, was also found 13 times in the WebCorp concordance lines. In those examples it was followed by *language* (9)*, human language*, *linguistics*, *linguistic data*, and *a body of electronic text*, once again illustrating widely differing interpretations. This pattern indicates the tendency to define corpus linguistics with reference to what researchers working with/in corpus linguistics *do* (which, of course, also suggests another, possibly more fruitful, way of addressing our initial question).

In addition to the explicit definitions, there were many 'working definitions', and the concordance lines in this section serve to illustrate the variety of interpretations of *corpus linguistics* in the CLRA-Y sub-corpus:

is view is that *corpus linguistics* should be defined as a **method** and not as a theoretical f
ways, the position of *corpus linguistics*, as a powerful **methodology-technology**, is well-
the unwanted uses is a familiar **methodological problem in** *corpus linguistics.*
discuss the **methodological** pros and cons of *corpus linguistics* as opposed to traditional
It is in these cases that we need to use **the methodology of** *corpus linguistics*.
is point (Section 3), **the methodology of** *corpus linguistics* is used to examine some be
applied fruitfully to corpora of all sizes, *corpus linguistics* **methodology** comes into its
explanation concerns both **the methodology** and ideology of *corpus linguistics*.
posture verbs, are best substantiated through **the methodology of** *corpus linguistics*.
exts can be segmented and annotated in the spirit of *corpus linguistics* **methodology**.
Only when **methods of** *corpus linguistics* are applied in areas of linguistic research can

(CLRA, 8 files)

This first group shows examples where corpus linguistics is associated with *method**, but clearly there are differences in the concordance lines, and not only in the choice of the term *method* or *methodology.* For example, references to the *methodology of corpus linguistics* do not specify whether corpus linguistics is viewed as a methodology or, for instance, a theoretical approach possessing a particular methodology. This pattern is also seen in the co-occurrences of *technology* and *tool*, where once again *corpus linguistics* may often be interpreted as 'possessing' rather than 'being':

Corpus linguistics has also established itself as an important **tool** for
Aside from a conviction that the new **technologies of** *corpus linguistics* have made
can all be investigated using **the tools of** *corpus linguistics*—indeed, it
a **key tool in** *corpus linguistics* is the concordancer which gives the researcher
in with concordancing as it is **a core tool** for analysis **in** *corpus linguistics*.

(CLRA, 5 files)

In the research articles, *corpus linguistics* is also defined as an *approach*, but as we have seen from the *WebCorp* concordance lines this can refer to either a theoretical or a methodological approach and so leaves its scientific status or nature uncertain:

The *corpus linguistics* **approach** in the study of health language has the potential
recent **approaches**, including *corpus linguistics* (Firth, 1957; Sinclair, 1987,
d systematically introduce students to *corpus linguistics* as **a means** to study 1. duction
Corpus linguistics, as a usage-based **approach** to the study of language,

(CLRA, 4 files)

Corpus linguistics is also represented as an 'enabler'. This may be seen as similar to the methodology view, but it additionally serves to offer a positive evaluation and so to justify the research reported in the articles. This favourable evaluation is emphasised in line 11 (marked with an asterisk) with the phrase *the advent of*, which also appeared in another article in the corpus with reference to *corpus linguistics*:[12]

	the **benefit of** this quantum leap in language awareness which	*corpus linguistics* **affords**.
	e the practical **benefits of**	*corpus linguistics* and data-driven learning, a number of
	e analysis has shown that	*corpus linguistics* can make **contributions** far beyond
	of data.	*Corpus linguistics* is probably most widely known for its **contributions** to
	d said (16), for example.	*Corpus linguistic*s may also **help** to explain the origin of
	scipline (Stubbs, 2001).	*Corpus linguistics* **leads to** a more 'evidence-based'
	ding large bodies of texts,	*corpus linguistics* **offers the possibility of unveiling**, for
	e overall development of	*corpus linguistics* has made **possible** into 'word-based'
	t the new technologies of	*corpus linguistics* have made it **possible** to purely observe
	possible **contribution** of	*corpus linguistics* and data driven learning to the field. The
	ption. With the advent of	*corpus linguistics*, it has become **possible** to discover*
	s. One student reflected, ''	*corpus linguistics* has **revealed** the inadequacies of
Corpus linguistics, as a usage-based approach to the study of language, **provides** l		
r students to share the insights into language use that *corpus linguistics* **provides**. In this		
assistance may come from interdisciplinary contacts with computer technology and *corpus linguistics*		

(CLRA, 9 files)

References to *corpus linguistics* as a *discipline* may be seen in the concordances below. Line 2 also illustrates the conceptualisation of *corpus linguistics* as possessing both theory and a method:

Corpus linguistics defines itself as a **discipline** the aim of which is to ''describe what is

The essentially social (rather than cognitive) orientation of *corpus linguistics* also surfaces in contemporary explorations of the **discipline's** theory and method.

(CLRA, 2 files)

The last group of concordance lines illustrates working definitions of *corpus linguistics* as a *field*:

rding to the perception of language within the **field** of	*corpus linguistics* sketched
l semantics and pragmatics coming from the **field** of	*corpus linguistics*, in
slation issue, particularly within the developing **field** of	*corpus linguistics*,
e useful to them. On the contrary, the growing **field** of	*corpus linguistics* offers much
ing in the **fields** of NLP, computational linguistics and	*corpus linguistics* are
argue that the two **fields** of health care research and	*corpus linguistics* can be

(CLRA, 5 files)

Although less explicit, the frequencies of *in* and *within* also indicate that *corpus linguistics* is perceived by the authors as a *field*, or as an academic community, rather than, for example, as a *tool*:

dancing as it is a core tool for analysis	**in** *corpus linguistics*. Concordancing is the pr
ell established tagger programs used	**in** *corpus linguistics* have been designed to h
fference between the use of computers	**in** *corpus linguistics* versus in all those meth
pus-based' research and corpus-driven	**in** *corpus linguistics*. The former is describe
(Simpson et al., 2002). This growth	**in** *corpus linguistics* has resulted in the devel
sed in different contexts. A key tool	**in** *corpus linguistics* is the concordancer whi
exis). 9 Following standard practice	**in** *corpus linguistics*, I am not giving individua
s is a familiar methodological problem	**in** *corpus linguistics*. Ideally, one would like
B15 2TT, UK Abstract Work	**in** *corpus linguistics* has led to the developm
ctional grammar and recent advances	**in** *corpus linguistics*. The general approach
d has been an important area of study	**in** *corpus linguistics* since the 1960s (Sinclai
nitially because they are widely used	**in** *corpus lingu*istics (see, for example, Barn
of Europe then carries what is known	**in** *corpus linguistics* as a negative semantic
ts, this time in the fight of recent work	**in** *corpus linguistics*, artificial intelligence an
would require advances to be made	**in** *corpus linguistics* too. It would entail goin
tation towards the role of ethnography	**in** *corpus linguistics* as they view corpus-ba
e one of the more important concepts	**in** *corpus linguistics*. However, while other c
dby Sinclair), and is by now a classic	**in** *corpus linguistics*. What Sinclair noted
l system, and that recent research	**within** *corpus linguistics* points in this direction,
s claim comes from recent research	**within** *corpus linguistic*s. 4. Corpus linguistics
as for decades been a central figure	**within** *corpus linguistics*. Like Firth, Sinclair ha
nal problem; as is recent research	**within** *corpus linguistics*, Levinson questioned

(CLRA, 13 files)

From these concordance lines, it emerges that the researchers in this corpus predominately view *corpus linguistics* as a field or as a community of practice they work in, and which 'offers' possibilities. The high frequency of conceptualisations of corpus linguistics as something which the researcher works *in/within* also offers a possible key to understanding the concentration of occurrences of *corpus linguistics* towards the beginning and end of the research articles. While *corpus linguistics* is often discussed in terms of *method*, *methodology*, and also *technology* and *tool*, whether it is viewed as possessing" these functions or "being" one of them is often unclear. Indeed, corpus linguistics is only defined once as clearly being a tool. This is particularly interesting as explicit debate on the nature of corpus linguistics has frequently focused on the rejection of contrasting definitions such as tool or theory. The notion of *corpus linguistics* as theory is also very limited in this corpus. Although the presence of clusters like *semantic prosody* (a key cluster for the sub-corpus CLRA-Y) may suggest that, for these practitioners, *corpus linguistics* has developed its own theories of language study, it is never described as "being" a theory in itself.

4 Conclusion

From these small-scale studies it appears that there are perceived "difficulties" with the term *corpus linguistics*, first of all because it is far less frequent than would be expected. This is most surprising in the results from the CLRA corpus where the articles had been selected because they used the term *corpus linguistics* in the keywords or, where keywords were absent, in the title or the abstract. There are, of course, many possible ways of interpreting this unexpected infrequency and here only a few suggestions can be offered. One explanation could be that the term itself is problematic, and/or it may be related to the availability of so many alternatives. As Léon (2005) notes, "what is called 'Corpus Linguistics' covers various heterogeneous fields ranging from lexicography, descriptive linguistics, applied linguistics – language teaching or Natural Language Processing – to domains where corpora are needed because introspection cannot be used, such as studies of language variation, dialect, register and style, or diachronic studies" (2005: 36). Alternatively, as corpus linguistics is still considered relatively new, perhaps only time will decide what appellation is finally adopted; we might note that terms such as *computer corpus linguistics* and *computer corpora* have already been abandoned as the involvement of the computer is taken for granted. In contrast, it may be that *corpus linguistics* is now so well established that there is no need to 'introduce' it within the main body of the research article. This notion would certainly be supported by the results from the CONF

corpus, where the target audience of the papers were corpus linguists. Or it may be that *corpus linguistics* is used in the keywords to mark the quantitative orientation or field of the work, and having established the nature of the study, the researcher then simply goes on to report the findings.

Second, within the research articles that do employ the term *corpus linguistics*, there are radical differences in the representation and understanding of what corpus linguistics is, as seen above. While such variation is interesting in itself, it should also be noted that the different descriptions of the status are not all necessarily incompatible. Several articles contained both references to *methodology* and *field*, and this may be explained by the transversal nature of corpus linguistics, and by the way in which the discipline is driven by the technology, as highlighted by Partington, who states that:

> To make such clear distinctions between instruments and enterprise is anachronistic. Firstly, because observation informs theory just as theory informs observation: […] Secondly, in most modern physical science, the object of observation is only tangible, in a sense only *exists* for an observer (outside a mathematical formula), through the instruments of study, which thus constrain not only what can be perceived but even the very questions that can usefully be put of the physical world.
>
> Partington (forthcoming)

Similarly, Mukherjee (2005) notes that the terms discipline and methodology are not mutually exclusive, stating that "I would contend that that corpus linguistics represents both a new method (in terms of computer-aided descriptive linguistics) and a new research discipline (in terms of a new approach to language description)" (2005: 86), and he draws an analogy with the introduction of the microscope, leading to the creation of the discipline of microbiology.

Furthermore, it may be that corpus linguistics is resilient to clear definition because, as Gries (2006b: 4) suggests, it "seems to be a category with a prototype structure: there are a few criteria that are – though not individually necessary – shared by much, if not most, work within corpus linguistics, and there is a variety of criteria which are less central to the work of many corpus linguists" (see also Williams 2006).

However, as stated at the beginning, the aim of this paper is not to try to provide any definitive answer to the opening question of "what is corpus linguistics?" but rather to simply bring it to the reader's attention once again, perhaps to encourage a little reflection (even if this goes no further than a quick concordance of our own papers!) and comparison. Apart from the intrinsic interest, it does seem that trying to describe the status of corpus linguistics, at least for our own individual operational purposes, may potentially lead to a greater awareness of the scientific context within which each of us works.

Notes

1. http://torvald.aksis.uib.no/corpora/1998-3/0006.html (retrieved 11 May 2007).
2. This keyword selection would have excluded many articles from the *International Journal of Corpus Linguistics*, but as most articles could not be converted to .txt this was not, in practice, an issue.
3. There was some subjectivity in the decision as to what was considered an introduction or a conclusion. Where possible, the criterion applied was what the authors had described as introduction or conclusion; in the absence of clear markers they were simply defined as the first and last sections of the article.
4. http://www.corpus.bham.ac.uk/PCLC/ (retrieved 11 May 2007).
5. Occurrences per thousand words.
6. Almost half of the occurrences in the bibliographies were made up of the following: *International Journal of Corpus Linguistics* (23). *Corpus Linguistics in North America* (15). *Corpus Linguistics: investigating language structure and use* (9). *Corpus Linguistics at Work* (8).
7. In order to reduce the effect of skewing, instances which occurred in fewer than five articles were not included.
8. Keywords and Key clusters which occurred in fewer than five articles were not included. Overall this meant that more keywords were excluded from the CLRA-N sub-corpus than the CLRA-Y sub-corpus as there appeared to be less common ground between the articles in the former.
9. *Firth* also appeared as keyword but was only present in four different articles.
10. http://www.webcorp.org.uk/ Described on its website as "a suite of tools which allows access to the World Wide Web as a corpus" (retrieved 11 May 2007).
11. Six lines were excluded which did not contain definitions of corpus linguistics (retrieved 11 May 2007).
12. In the BNC, in addition to references to innovation (*new, modern*), the phrase *the advent of* seems to show a semantic preference for the field of technology (*television, computer*). Findings were similar in a large corpus of newspapers from 2005, although in this case collocates also included *internet, digital* and *broadband*.

References

Aarts, Jan. 2002. Does corpus linguistics exist? Some old and new issues. In L. E. Breivik and A. Hasselgren (eds.). *From the COLT's mouth… and others': Language corpora studies in honour of Anna-Brita Stenström*, 1–19. Amsterdam: Rodopi.

Aarts, Jan and Willem Meijs (eds.). 1984. *Corpus linguistics: Recent developments in the use of computer corpora in English language research*. Amsterdam: Rodopi.

Aarts, Jan and Theo van den Heuvel. 1982. Grammars and intuitions in corpus linguistics. In S. Johansson (ed.). *Computer corpora in English language research*, 66–84. Bergen: Norwegian Computing Centre for Humanities.

Bowker, Lynne and Jennifer Pearson. 2002. *Working with specialized language: A practical guide to using corpora*. London: Routledge.

Carter, Ronald. 2004. Introduction to J. Sinclair and R. Carter (ed.). *Trust the text: Language, corpus and discourse*, 1–6. London: Routledge.

Chomsky, Noam. 2004. (Interviewed by Andor, Jozsef). The master and his performance: An interview with Noam Chomsky. *Intercultural Pragmatics* 1(1): 93–111.

Fillmore, Charles. 1992. Corpus linguistics or computer-aided armchair linguistics. In J. Svartvik (ed.). 35–60.

Gries, Stefan Th. 2006a. Some proposals towards more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik* 54(2): 191–202.

Gries, Stefan Th. 2006b. Introduction to S. Gries and A. Stefanowitsch (eds.). *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis*, 1–18. Berlin, Heidelberg, New York: Mouton de Gruyter.

Hoey, Michael. 1993. Introduction to M. Hoey (ed.). *Data, description, discourse: Papers on the English language in honour of John McH Sinclair*, v-ix. London: Harpercollins.

Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. In J. Svartvik (ed.). 105–122.

Léon, Jacqueline. 2005. Claimed and unclaimed sources of corpus linguistics. *Henry Sweet Society Bulletin* 44: 36–50.

Mahlberg, Michaela. 2005. *English general nouns: A corpus theoretical approach*. Amsterdam: John Benjamins.

Mahlberg, Michaela. 2006. Lexical cohesion: Corpus linguistic theory and its application in English language teaching. *International Journal of Corpus Linguistics* 11(3): 363–383.

McCarthy, Michael. 2001. *Issues in applied linguistics*. Cambridge: Cambridge University Press.

McEnery, Tony, Richard Z. Xiao and Yukio Tono. 2005. *Corpus-based language studies: An advanced resource book* . London: Routledge.

McEnery, Tony and Andrew Wilson. 1996. *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McEnery, Tony and Costas Gabrielatos. 2006. English corpus linguistics. In B. Aarts and A. McMahon (eds.). *The handbook of English linguistics*, 33–71. Oxford: Blackwell.

Meyer, Charles F. 2002. *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.

Morley, John and Paul Bayley (eds.). Forthcoming. *Wordings of war. Corpus assisted discourse studies on the war in Iraq*. London: Routledge.

Mukherjee, Joybrato. 2005. *English ditransitive verbs: Aspects of theory, description and a usage-based model.* Amsterdam and New York: Rodopi.

Partington, Alan. 2004. Corpora and discourse: A most congruous beast. In A. Partington, J. Morley and L. Haarman (eds.). *Corpora and discourse*, 1–20. Bern: Peter Lang.

Partington, Alan. Forthcoming. Evaluating evaluation and some concluding thoughts on CADS. In J. Morley and P. Bayley (eds.).

Römer, Ute. 2005. *Progressives, patterns, pedagogy. A corpus-driven approach to English progressive forms, functions, contexts and didactics*. Amsterdam: John Benjamins.

Scott, Mike. 1999. *WordSmith Tools version 3.* Oxford: Oxford University Press.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stubbs, Michael. 1993. British traditions in text analysis: From Firth to Sinclair. In M. Baker, F. Francis and E. Tognini-Bonelli (eds.). *Text and technology: In honour of John Sinclair*, 1–36. Amsterdam: John Benjamins.

Stubbs, Michael. 2001. *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

Stubbs, Michael. 2006. Corpus analysis: The state of the art and three types of unanswered questions. In G. Thompson and S. Hunston (eds.). 15–36.

Svartvik, Jan (ed.). 1992. *Directions in corpus linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991*. Berlin and New York: Mouton de Gruyter.

Teubert, Wolfgang. 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics* 10(1): 1–13.

Teubert, Wolfgang and Ramesh Krishnamurthy. 2007. General introduction to W. Teubert and R. Krishnamurthy (eds.). *Corpus linguistics. Critical concepts in linguistics*, 1–37. London and New York: Routledge.

Thompson, Geoffrey and Susan Hunston (eds.). 2006. *System and corpus: Exploring connection*s. London: Equinox.

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam: John Benjamins.

Williams, Geoffrey. 2006. La linguistique de corpus: une affaire prépositionnelle. In F. Rastier and M. Ballabriga (eds.). *Corpus en lettres et sciences sociales: des documents numériques à l'interprétation. Actes du colloque international d'Albi, juillet 2006*, 151–158. Paris: Texto.

Appendix. Journals from which articles forming the CLRA corpus were taken

Computers and Composition
Critical Discourse Studies
Discourse & Society
Discourse Studies
English for Specific Purposes
International Journal of Corpus Linguistics
Journal of Applied Linguistics
Journal of Biomedical Informatics
Journal of English for Academic Purposes
Journal of Pragmatics
Journal of Second Language Writing
Language in Society
Language Sciences
Linguistics and Education
Literacy
Nordic Journal of Linguistics
Pergamon Language Sciences
System
Target
TESOL Quarterly