Pre-publication copy. Final version published as

Michael Stubbs (2001) Texts, corpora and problems of interpretation: A response to Widdowson. *Applied Linguistics*, 22, 2: 149-72.

TEXTS, CORPORA AND PROBLEMS OF INTERPRETATION: A RESPONSE TO WIDDOWSON

Michael Stubbs

ABSTRACT

Widdowson (2000) criticizes two approaches to language description – corpus linguistics and critical discourse analysis – which both concentrate on 'real' (i.e. attested) language. His main criticism focuses on work which combines these two approaches by attempting to use corpus data in order to remedy deficits in critical discourse analysis. He raises important points about text interpretation, and therefore about the relation between corpus linguistics and social theory. However, his argument is flawed by its misrepresentation of the data, methods and central concepts of corpus linguistics. In particular, he ignores the logic involved in comparative analyses of variable and quantitative corpus data.*

1. INTRODUCTION

In a recent edition of *Applied Linguistics*, Widdowson (2000, henceforth W) criticizes work in corpus linguistics and in critical discourse analysis. In particular, he argues that descriptions based on corpus evidence are only partial and warns against their application in text interpretation. His article fits into other work in which he is sceptical of politically committed approaches to stylistic interpretation (Widdowson 1992, especially vii-xiv, 182ff), and into a series of articles in which he has repeatedly criticized, sometimes very severely, corpus linguistics (Widdowson 1991) and critical discourse analysis (Widdowson 1995a, b, 1996a).

Although W raises important issues about text interpretation, I will argue that he touches only the tip of the interpretative iceberg: since he does not discuss the inherently quantitative, variable and comparative nature of corpus data, he cannot directly discuss the relations between textual, cognitive and social phenomena.

In fact, W criticizes three distinct areas: (1) corpus linguistics, (2) critical discourse analysis, and (3) the applications of corpus data in language teaching. I will deal only with (1) in any detail, since here lie the fundamental questions. Only if we can be clear about the nature of corpus data and methods can wider questions of their applications in language education be tackled. I discuss (2) only in passing, and (3) not at all. [NOTE 1.]

2. THE BACKGROUND ARGUMENTS: DESCRIPTION AND APPLICATIONS

There are two background arguments in W. The first is a long-running debate, to which Widdowson himself contributed so influentially, namely: can concepts from theoretical linguistics be applied directly to real world problems (the 'linguistics applied' position), or must applied linguistics develop its own theories, which mediate and interpret findings both from linguistics and also from other disciplines (the 'applied linguistics' position)? Nowadays, partly thanks to Widdowson, the second position may seem self-evident, though whether we need a separate layer of mediation is doubtful. In line with the second position, W questions whether descriptions of language use, especially those based on corpora, can be applied to textual interpretation.

The second argument is that, since the 1980s, linguistics has undergone a profound shift from a primary interest in internalized I-language to externalized E-language (Chomsky 1988), that is from introspective to attested data. Two developments have led to this increased interest in 'real' language: the technology which now allows corpus linguists to describe very large quantities of text; and the attempt by critical discourse analysts to reveal the ideological assumptions of texts. W argues that neither perspective, on I-language or E-language, provides the whole truth. Presumably, therefore, they should be combined, though W makes no proposal as to how this might be done.

W accepts that corpus linguistics is 'an immensely important development in descriptive linguistics', which has revealed a previously unsuspected 'reality about language usage' (p.6), but he emphasizes that this provides 'only a partial account of real language' (p.7). The partiality is evident, he argues, in the lack of correspondence between corpus findings and native speaker intuitions: since they are contrary to intuition, they cannot be the full story.

So, the problems concern the relations between linguistic descriptions, the unsuspected reality which they reveal, and interpretations of these descriptions. And this involves very different things: interpretations are subjective, but they must nevertheless be related to findings which are objective, in so far as they have been discovered by replicable methods in publicly accessible data. In the context of critical discourse analysis, this leads us into deep Whorfian waters, when patterns of language use are related to ideologies held by individuals or social groups (Stubbs 1997b). We must try to disentangle public data and private interpretations, cause and correlation, and also weak and strong forms of Whorfian arguments. For example, it might be that systematic differences in language use correlate with, but do not cause, identifiable ideologies. Everything therefore depends on whether we can provide a clear statement of the logic of the positions.

3. THE DATA AND METHODS OF CORPUS LINGUISTICS

First therefore, we require an accurate statement of the data and methods of corpus linguistics.

3.1. Possible, attested and probable.

W (p.7) follows Hymes (1972) in distinguishing between what is formally possible, contextually appropriate and actually attested, and claims (p.7) that corpus linguistics deals only with the textually attested. He then repeatedly opposes 'the attested' and 'the possible' (pp.7-8, 23, but also pp.10, 19). The misleading nature of this opposition becomes most apparent perhaps in this statement (p.8): 'it would be [...] mistaken to suppose that what is textually attested uniquely represents real language'.

But who supposes this? Not, as far as I am aware, any corpus linguists. Corpus linguistics is not concerned with what happens to occur (at least once): indeed its methods are generally designed to exclude unique instances, which can have no statistical significance. It is concerned with a much deeper notion: what frequently and typically occurs. What frequently occurs in texts is only a small proportion of what seems to be possible in the system (Pawley & Syder 1983), and the more relevant opposition is between what is possible and what is probable (Kennedy 1992).

In any case, instances can be interpreted only against a background of what is typical. Corpus linguistics therefore investigates relations between frequency and typicality, and instance and norm. It aims at a theory of the typical, on the grounds that this has to be the basis of interpreting what is attested but unusual. Priority is given to describing the commonest uses of the commonest words. (Sinclair et al 1998 illustrate software which gives an operational definition of typicality.)

W's repeated use of the term 'attested' subtly colours his whole argument. It is important to be clear whether any given data fragment has actually occurred, or whether it has been invented by the linguist as an illustration. But any single occurrence is, in itself, of little interest for the description of the language as a whole. [NOTE 2.]

3.2. Observational data, introspective data and mental models.

W distinguishes (p.6) three complementary types of data: third-person observations, second-person elicitations and first-person intuitions. What do *they* actually say? What would *you* say? And what do *I* think *I* say? (See also Widdowson 1996b: 72-73.) This is a valuable and elegant suggestion, but W does not discuss how these three levels of reality relate to each other, or how such relations could be empirically investigated. [NOTE 3.]

Long before corpus linguistics, we knew that people do not talk as they believe they do, and corpus linguists now often point out how radically intuition and use may diverge. Certainly, these relations between behavioural and psycholinguistic data are under-investigated, but a start has been made. Fillmore (1992) provides a detailed argument for combining corpus-based and introspective data; Moon (1998) uses corpus data to propose lexical schemas and prototypes; and Sinclair (1991a: 113) proposes a specific hypothesis about the systematic relation between intuition and use. In order to answer questions such as 'what is the meaning of a given linguistic form?', we have to study quantitative data on its uses, admit the variability of the examples, and formulate a prototype.

3.3. Partiality, point of view and reality.

W argues that 'the linguistics of the attested is just as partial as the linguistics of the possible' (p.7, also pp.3, 5, 24), but admits that 'all enquiry is partial' (p.23). He is also sceptical of attempts to study language in the 'real' world (pp.3, 5), yet he concedes that corpus analysis reveals 'a reality about language usage which was hitherto not evident to users' (p.6). Burrows (1987: 2-3) elegantly formulates the paradoxical nature of this reality:

> 'Computer-based concordances, supported by statistical analysis, now make it possible to enter hitherto inaccessible regions of the language [which] defy the most accurate memory and the finest powers of discrimination.'

So, what is it that we can see from this new point of view? A set of concordance lines is a sample of a node word together with a sample of its linguistic environments, often defined as a span of words to left and right. In Saussurean (1916: 171) terms, a syntagmatic relation holds between items *in praesentia*, which co-occur in a linear string. A concordance line is a fragment of *parole*, where a single instance of syntagmatic relations can be observed. We are interested in more, however, than what happens to have occurred once in such a fragment. A paradigmatic relation is a potential relation between items *in absentia*, which have a psychological reality ('des termes *in absentia* dans une série mnémonique virtuelle', p.171). If paradigmatic relations are seen as a virtual mental phenomenon, then they are unobservable.

In an individual text, neither repeated syntagmatic relations, nor any paradigmatic relations at all, are observable. However, a concordance makes it possible to observe repeated events: it makes visible, at the same time, what frequently co-occurs syntagmatically, and how much constraint there is on the paradigmatic choices. The co-occurrences are visible on the horizontal (syntagmatic) axis of the individual concordance lines. The repeated paradigmatic choices – what frequently recurs – are equally visible on the vertical axis: especially if concordance lines are re-ordered alphabetically to left or right. (Tognini- Bonelli 1996.)

Since concordances make repetitions visible, this can lead to an emphasis on the repetitive and routine nature of language use, possibly at the cost of striking individual occurrences (the difficult relation between frequency and salience again). Frequency is not necessarily the same as interpretative significance: an occurrence might be significant in a text precisely because it is rare in a corpus. But unexpectedness is recognizable only against the norm.

These repetitions can now be studied. A major part of the patterning revealed by concordances is the extent of phraseology, which is not obvious to speakers, and has indeed been ignored by many linguists. The patterns have been discovered, but not created, by the computer. The test of this claim, and a major strength of computer-assisted corpus analysis, is that findings can be replicated on publicly accessible data: there is always an implicit prediction that you will find the same patterns in independent corpora. These probabilistic semantic patterns (collocations, colligations, etc) revealed across many speakers' usage in corpora are not within the control of individual speakers, and are not reducible to anything else (Carter & Sealey 2000). Where I agree with W is in his insistence that their cognitive influence has yet to be stated clearly.

3.4. Interpretation and convention.

W emphasizes the different possible interpretations of lexical and grammatical features.

However, one of the deepest problems – which W does not raise – is the relation between interpretation and convention. It is currently fashionable to emphasize the interpretative aspects of text analysis, and to play down the pervasive patterning in data, and many theorists are sceptical of the view that meanings are explicit in text. This scepticism is evident both in linguistic theories of pragmatics, such as relevance theory (Sperber & Wilson 1995), and also in a broad tradition of interpretative sociology (to which W, p.6, alludes), in work by Garfinkel and Cicourel onwards.

Batstone (1995), also with reference to critical discourse analysis, tries to distinguish between stable semantic (notional) aspects of textual meaning and unstable context-dependent pragmatic (attitudinal) meaning. However, Levinson (1983: 11) points out that some pragmatic meanings are conventionally encoded. And a major finding of corpus linguistics is that pragmatic meanings, including evaluative connotations, are more frequently conventionally encoded than is often realized (Kay 1995, Moon 1998, Channell 2000). Both convention and interpretation are involved, but it is an empirical question to decide how much meaning is expressed by conventional form-meaning relations, and how much has to be inferred.

Concepts of convention and norm raise problems in the not infrequent cases when interpretations diverge. I have no space here for detailed examples, but readers might check the divergent connotations given for *cronies* in different corpus-based dictionaries. Is it a neutral word for "(male?) friends"? Or a pejorative word connoting "disreputable friends"? Or does it even imply "criminal activities"? These divergences are themselves open to empirical corpus study.

3.5. Process and product.

W repeatedly argues that corpus linguistics provides us with a description of text as product, not discourse as process (pp.6, 9, 10). Since a text is a 'static semantic patchwork' (pp.7, 17, 22), which has been taken out of its social context of inference and interpretation, we can study only 'textual traces' (pp.7, 11, 21, 22) of discourse process.

This is perfectly true, though the problem is very widespread in empirical disciplines. Recognizing the problem obviously does not solve it, but it shows that corpus linguistics is trying to develop observational, empirical methods of studying meaning, which are open to the same tests as are applied in other disciplines. For example, consider the parallels between corpus linguistics and geology, which both assume a relation between process and product. By and large, the processes are invisible, and must be inferred from the products.

Geologists are interested in processes which are not directly observable, because they take place across vast periods of time. What is observable is individual rocks and geographical formations: these products are the observable traces of processes which have often taken place a long time in the past. They are highly variable, because any specific instance is due to the local environment. Nevertheless, these variable products are due to highly general processes of destruction (such as erosion) and construction (such as sedimentation). (Love 1991.)

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. What is directly observable is the individual products, such as utterances and word combinations. (In addition, repetitions of such patterns, across time, can be made observable if different occurrences are displayed by concordancers and other software: see above.) These individual word combinations are the observable traces of general patterns of collocation and colligation. They are highly variable due to local sociolinguistic contexts. Nevertheless, these variable products are due to highly general processes of probability and speaker expectation.

3.6. Summary.

W's account of corpus linguistics, and hence of associated problems of interpretation, lacks a discussion of

> the empirical, observational methods used in corpus semantics
> the ontological status of the patterns which are revealed
> the balance in language use of convention and interpretation
> the relation between individual instances and general patterns.

4. PRESENTATION CONVENTIONS

In the remainder of the article, all examples are attested [A] in corpus data, unless explicitly marked as invented [I], or modified [M]. LEMMAS (LEXEMES) are in upper case. *Word-forms* are lower case italic. "Meanings" are in double quotes. 'Quotes' from other authors are in single quotes.

5. THE SPECIFIC ARGUMENTS: CO-SELECTION AND CONTEXT

Although W's main argument is with corpus linguistics, his few examples do not draw on quantitative data. On the contrary, he uses individual example sentences, in order to argue that the meaning of their grammatical features (such as transitive and intransitive) cannot be read off their encoding in an individual text. His arguments do not therefore tackle central interpretative problems which arise with quantitative data.

Corpus study is predicated on concepts of co-selection and co-occurrence, so, with some of his examples of the importance of phraseology, W is pushing at an open door. For example, he discusses the lemma CLOSE (p.20), which is interpreted differently in different contexts (physical and phraseological) such as:

[1] Closed (on a sign in a shop door)
[2] a closed shop (in the sense of "employment closed to those who are not members of a trade union")

Corpus study also shows that different word-forms of a lemma frequently have different collocates and different senses: for example, in the "closed shop" sense, only the *ed*-form can occur. Further senses of CLOSE occur in phrases such as

[3] a closed shop; a closed mind; the shop was closed; the road was closed; a closing down sale; the closing stanza; close the door; close the discussion

Phraseology often stretches further than just one or two words on each side of the node. Using invented examples, W (p.20) argues that [4] (= W example 7) might refer to industrial decline, whereas [5] (= W example 9) is more likely to refer to the end of the day's business.

[4] industrial premises and shops were closing [I]
[5] the shops in Oxford Street close at six [I]

W's point is that linguistic features never occur alone, though he does not identify explicitly which features lead to which interpretation in this case. I hypothesize that it is reference to time of day, helped in [5] by simple present tense, that would normally be interpreted as: "the shops usually close for the night at six". Therefore, with a little ingenuity, and invented data, we can come up with examples such as

[6] industrial premises and shops were closing, the evening sky was darkening, and people were hurrying home after work [I]

The point that meaning occurs as a prosody, due to the combination of different linguistic features across phrases of indefinite extent in text, is massively corroborated by corpus studies (such as Sinclair 1991a, Louw 1993). However, introspective ingenuity applied to invented sentences tells us nothing about what usually occurs.

In addition, W's arguments are based on two kinds of circularity. First, invented examples are not independent of the analyst, and the theory is not independent of its supporting data. A theoretical point is proposed, and an example is invented to support it. Second, W argues for the possibility of different interpretations. But if I provide alternative interpretations of his examples (as I have done), then this confirms his argument. His argument is unfalsifiable since no empirical evidence is relevant to the dilemma of interpretation which it poses.

5.1. Artificial ambiguities.

A further reason to be sceptical of ingenious invented examples is that they can greatly exaggerate the ambiguity of language in use. Speakers can handle cases such as [4] versus [6], but they rarely have to, since it is rare to find different senses of a lemma or word-form in the same text. To take a clichéd example, BANK (= "financial institution" versus "raised land at the edge of a river, under shallow water, etc") is obviously ambiguous on its own, but then it never occurs on its own (except in linguistics textbooks). As Saussure taught us, words have meaning only in relation to other words. In a small corpus of a few million words, I was unable to find a text in which the lemma is used in both the "money" and the "water" senses. In addition, the intended sense was often signalled in several independent ways, by a fixed phrase and also by several different collocates in lexical sequences such as

[7] money – deposits – Bank of England – paid – instalment
[8] shallows – sea – cod – Icelandic Banks – haddock

5.2. Context.

W accuses corpus methods of ignoring context. He argues that we cannot infer the significance of 'features in isolation' (p.19), that corpus data are decontextualized, and that we find 'an analysis of text which is then given unwarranted significance in disregard of crucial contextual factors' (p.22). (His own examples are nevertheless largely invented, and therefore decontextualized.) W's view of context emphasizes the 'ethnographic' or 'sociocultural settings' (p.22) of texts. This view of context is perfectly valid, but does not invalidate alternative concepts.

To accuse corpus linguistics of ignoring context is strange, since it is essentially a theory of context: the essential tool is the concordance, where words are always studied in their contexts. A concordance seems to imply that context means just a few words of co-text. However, first, this also implies a hypothesis about how much context is relevant to establishing meaning. This is an empirical question, and at the level of micro context, it has produced the surprising finding that a short span of a few words to left and right is often enough to disambiguate words or to identify their evaluative connotations (Sinclair 1991a, Stubbs 1995, Clear 1996). If critics are uncomfortable with this concept of context, then it is up to them to show that more context is necessary. [NOTE 4.] Second, as we saw above, on their vertical axes concordances also show repetitions and therefore inter-textual relations. In corpus work, context means two rather different things: not only co-text (a short span of a few words within one single text), but also inter-text (repeated occurrences, often a very large number, of similar patterns across different, independent texts). Its methods show how co-text can be systematically used to provide observational evidence of meaning, and its slogan is 'meaning is use'.

5.3. Individual utterances and frequent patterns.

Examples of individual utterances cannot tackle claims about the ideological implications of textual patterns. Consider a type of example which occurs in work in critical discourse analysis. When political events in Third World countries are reported in the press, two patterns are often present. First, people are often represented in large numbers (Said 1978: 287), and second, their actions are often described metaphorically, as in these two examples from Lee (1992):

[9] the black township [...] erupted
[10] the marchers [...] swept through a roadblock

If we put these two observations together, such utterances seem to imply that people act in large groups, as though they were some kind of natural force, like a volcano or a river in flood, that they have 'no individuality, no personal characteristics', and that their behaviour can be seen only as 'irrational' (Said 1978: 287).

Now, whether this interpretation is correct or not, presumably a single utterance is unimportant. Presumably, indeed, a single newspaper article is unimportant in the greater scheme of things. This is particularly so, since people's recall of news stories is very low indeed: rarely more than 30 per cent even immediately after a news broadcast, and sometimes as low as 5 per cent (Bell 1991: 232). (Anyway, perhaps metaphors of violence 'erupting' really are dead.) However, if such descriptions are regularly used in a wide range of reports, then they might come to seem a natural way of talking about things, and it is plausible that they come to influence how we think about such events. It is plausible, but the problem is how to prove it: there is always a category shift when we move from ways of talking to ways of thinking.

So, how does frequency affect interpretation? Antonius says: 'Brutus is an honourable man'. The words mean one thing, Antonius means something different, and the reader's interpretation changes as Antonius repeats the words. Advertisers certainly believe that repetition of a message influences behaviour. The relation between frequency, routine, convention and interpretation is difficult to document (though see Krishnamurthy 1996 for a good attempt using corpus data).

5.4. Summary.

I agree with W that interpretations and patterns of language use are quite different kinds of object. They imply, respectively, agency and structure, they exist on different time scales, and they are not reducible one to the other. Interpretations depend on individual human agency and are produced at a particular point in time. The patterns identified by corpus methods are a structural feature of language in use, produced over a long time period by many different speakers, and independent of the individual analyst. Since the patterns are probabilistic, they are not observable in single instances. They are features of a social *langue* (in a Saussurean sense), and the relation of this to individual competence (in a Chomskyan sense) is a difficult and unsolved problem. (On ontological questions which this distinction raises, see Carter & Sealey 2000.)

6. UTTERANCE AND CO-TEXT

The term 'context' means many different things. One meaning is position in a textual sequence, which again casts doubt on any method of analysis which relies on invented and isolated sentences. Let us try to make the argument here as explicit as possible.

An often used example concerns actives, passives and ergatives, as in (with apologies to Lakoff, Ross, Jackendoff and others) these invented sentences:

[11] Floyd broke the glass
[12] The glass was broken by Floyd
[13] The glass was broken when I came in
[14] The glass was broken
[15] The glass got broken
[16] The glass broke
[17] The breaking of the glass (could be heard a mile away)

Active transitive clauses in English must have a noun phrase as a subject/agent; passives may optionally have an explicit agent in a *by*-phrase; and ergatives, grammatically, cannot express an agent.

However, there are many reasons for omitting the agent. The speaker might wish to be vague or ambiguous (e.g. between stative and dynamic meanings). The agent might be omitted in order to avoid attributing blame (*it just broke*), and, in turn, this might be for legal reasons in a newspaper report (*the officer was repeatedly kicked in the head*: Biber et al 1999: 477). But it might also be because the information has been mentioned earlier in the text, or is unimportant or obvious (as in *he was arrested last year*: presumably he was arrested *by the police*). Or it may have reasons to do with cohesion and information flow, placing the focus on Floyd, the glass, the whole event, and so on.

Nominalization allows other information to be omitted, since a noun phrase does not mark tense, but again noun phrases have many functions. It is well known that passivized and nominalized styles are common in formal scientific writing, but this is not just because scientists like to think abstractly, and regard the world (in an inhuman way) as things and products rather than as events and processes. (On the functions of nominalization in scientific writing, see Halliday and Martin 1993, and Atkinson 1999.) For example, Halliday (1993: 55-6, 69) shows that complex noun phrases can have quite specific textual motivations: they can be used to refer to complex phenomena, and this affects the sequence in which forms are used in an individual text. In a text on how stress in glass causes it to crack, the following phrases occur, in this sequence:

[18] glass cracks ... a crack grows ... the rate at which cracks grow ... the rate of crack growth ... the glass crack growth rate

Similarly, in a popular science article about astronomy (*New Scientist*, 18 August 1990), I found that nominal forms were first used in the headline of the article (*galaxy evolution*) and the opening summarizing paragraph (*star formation*), presumably because they are shorter (yet another function of nominalization). Thereafter the verbal-nominal sequence follows the pattern identified by Halliday, as illustrated in [19] to [25]. Several propositions are first encoded as noun phrase plus verb, then later in the text the verbal content is integrated into a noun phrase:

[19] stars are born ... star formation
[20] the galaxy ages ... the age of its star clusters
[21] the metal ratio of the gas has changed ... the change in the metal ratio
[22] enriching the gas with metals ... metal enrichment

In this way, whole events can be encoded in still more complex noun phrases, such as

[23] the rate of star formation in a galaxy
[24] the rate of metal enrichment

which can be used as the subject or object of other verbs. Once the concept of speed of formation has been encoded in a noun phrase, then different rates can be compared with each other:

[25] *the change in the metal ratio* over time is a pretty good indicator of *the rate of star formation*

We now have the well-known grammar of such texts, with long complex noun phrases in simple clause structures: *NP is NP*.

In summary, no conclusion whatsoever can be drawn about the ideological function of an individual grammatical form, or even of a whole sequence, independently of textual organization. Variation may be due to the assumed knowledge of the intended addressee, place in a textual sequence [NOTE 5.], or text-type (this may have conceptual motivations, but may become conventional over time). A text should not be treated as a resource for psycho-social inferences as if it had no organization of its own. If we go directly from linguistic categories to psycho-social categories, we by-pass a layer of textual organization. We must therefore take into account at least three levels of description which are not reducible to each other:

> individual linguistic features (e.g. passives versus actives)
> their function in a textual sequence
> their cognitive or social function.

7. A QUANTITATIVE AND COMPARATIVE STUDY

W's main examples are based on a study (Stubbs 1994, revised as a chapter in Stubbs 1996), which analyses linguistic features in two school books. W's reason for discussing this study (pp.11-17, 19-21) is that it tries to combine the two developments of which he is sceptical, by applying the methods of corpus linguistics to critical discourse analysis. [NOTE 6.] W criticizes this study for moving too easily from formal features of the texts to interpretation. However, I will argue that he gives a highly partial account of the study, since he ignores its comparative and quantitative design. [NOTE 7.]

7.1. Interpreting comparisons.

First, W claims (p.11) that only 'one particular grammatical feature' is analysed in the study, namely ergativity. Actually, two sets of features are analysed. Now, two is not much more than one. However, the two sets of features are independent of each other, and in both cases, there are significant differences between their distributions in the two books. For the first set, 430 verbs were identified which can occur in three forms, transitive, passive and ergative. One such verb is EXPAND:

[26] Brazil has expanded its steel production (transitive)
[27] the refinery was expanded in 1981 (passive)
[28] Britain's cities have expanded outwards (ergative)

(Examples from Stubbs 1996: 137.) For all the verbs which occur in one or both books, the relative frequency of the three forms is compared across the two books.

In addition, the study compares: the five ergative verbs which are most frequent in both books, and which also occur in all three forms in one or both books (Stubbs 1996: 138-39); two individual verbs, since 'every verb has different syntax' (1996:

140); the frequency of passives in the texts, independently of ergative verbs (1996: 140); and the number of passives with *by* plus agent, and the number of human and abstract agents (1996: 140-41). W does not mention any of these comparisons.

The second set of linguistic features involves projecting clauses, which can either make explicit, or leave unattributed, the source of a proposition, as in, respectively:

[29] opponents of nuclear power say that [proposition]
[30] it has been predicted that [proposition]

Again the frequency of the forms is compared across the two books. And again, further comparisons are also made, between projecting clauses with personal and impersonal subjects (Stubbs 1996: 151).

The study presents further limited comparisons between the frequency of features in the two texts and in reference corpora. As software becomes available which can compare texts with corpora (e.g. Scott 1997), this will become an increasingly important type of analysis in future, and will raise further difficult problems of comparison between:

> the instance (an individual sentence or individual text)
> the norm for the text-type
> the norm for the language (as represented by a large general corpus).

7.2. Interpreting correlations.

Second, W argues (p.15): 'The only fact we have is that certain formal features occur with a certain frequency'. This is also inaccurate. We have two independent sets of features, whose exponents are differently distributed in two texts (with high levels of statistical significance). The main fact (finding) is therefore a correlation: between the frequency of linguistic features and the attitudinal stance of the authors (explicitly persuasive and politically committed in one case, implicitly neutral in the other). Now these differences may be misinterpreted. But the 'fact' is the correlation in two independent cases, and the probability of this occurring by chance is very small. (Indeed, as noted, there are more than two cases.) W does not discuss these quantitative or statistical data.

So, the most serious problem of interpretation – perhaps more serious than any identified by W – may be how to interpret correlations. This involves the problems of interpreting variable data, of inferring cause from correlation, of multiple causation, and so on. As is emphasized in the study itself (Stubbs 1996: 144):

> 'A [...] serious problem is that such stylistic patterns are probabilistic. There is no absolute difference between the two texts, and stylistic interpretation of frequency and probability data is very uncertain.'

8. ON EMPIRICAL SEMANTICS

So, W's view that the study selects particular grammatical features in order to 'identif[y] ideological stance' (p.12) turns the argument on its head. The ideological stance is given. To repeat: the finding is that this correlates with observable linguistic features.

There are two texts. It is known that one is explicitly politically committed (to an ecological stance), whereas the other is silent on such topics. It is hypothesized (because of what is known about text-types) that certain linguistic features will distinguish the two texts. This hypothesis is tested and strongly corroborated. There is no absolute difference: both texts use both possibilities in both sets of constructions, but the differences are statistically significant in both cases ($p < 0.001$ and $p < 0.01$). Other subsidiary patterns differ in comparable ways.

It is this argument structure which requires to be assessed. The analysis is not falsified by W, since he does not assess its findings or its internal logic. (Nor is it confirmed by my discussion here: indeed I have, in some ways, provided more far-reaching criticisms of my own study than W does.) The findings require to be tested by trying to replicate them on comparable but independent texts, and the logic remains to be tested, particularly to assess the claimed relations between patterns of language use and cognition.

Stubbs (1994/1996) and Widdowson (2000) are pulling in different directions. I emphasize descriptive methods and patterns, thereby possibly riding roughshod over the interpretation of individual utterances. W emphasizes the interpretations of individual sentences, thereby ignoring the data and methods which I present. I hold the view (in common with Widdowson) that we cannot arrive at a definitive interpretation of an individual text. I also hold the view (again, I assume, in common with Widdowson) that, as Firth put it, a statement of meaning 'cannot be achieved at one fell swoop by one analysis at one level' (1950: 192); meaning can be handled only by 'dispersing it in a range of techniques working at a series of levels' (1957: 7). However, I think it worthwhile to try and identify general mechanisms which contribute to textual meaning, and this involves interpreting individual utterances in relation to

> their place in a specific text sequence
> the norm in the genre (such as school textbooks)
> the norm in a wide range of text-types (as sampled in a corpus).

We must directly assess the theory of semantics which underlies corpus linguistics. The theory (traceable back to Wittgenstein, Firth and Austin) is that meaning is use, and this theory implies an empirical method: namely, observing which collocates frequently co-occur with a target word or other structure, and drawing inferences from this. This does not get us outside the circle of language: we can still only express meanings in words. But it avoids a unique reliance on the analyst's own words, it suggests specific hypotheses about the units of meaning (not words, but larger lexico-grammatical schemas), it shows the limitations of a compositional theory of semantics, and so on. And it uses publicly accessible observational data to study meanings, and therefore makes statements which can be replicated and corroborated – or of course questioned and refuted – on independent data.

For short, the theory and method can be called corpus semantics: the use of corpus evidence to study meanings (Teubert 1999). The methodology, which has been successfully used in writing major dictionaries and grammars, is set out by Hunston and Francis (2000). And Atkinson (1999) provides an impressive study, of precisely the kind which W wants, which combines linguistic and socio-cultural analysis: it compares text samples diachronically and puts them in their historical context.

9. INSTANCE AND NORM: A LEXICO-GRAMMATICAL EXAMPLE

Corpus semantics compares samples of language use with each other, whether two individual texts, a text and a reference corpus, or two constructions in the language. I cannot here illustrate all possible comparisons. However, my final criticism of W's argument is that it is not general enough. Let us therefore leave W's particular case and generalize the argument, by looking at comparative quantitative data on the meanings of passive constructions with BE and GET.

GET is one of the most frequent verbs in (especially informal spoken) English, and its semantics and pragmatics are complex. In an excellent corpus study, Johansson and Oksefjell (1996) discuss its functions as a quasi-auxiliary verb, and its uses in signalling "change" and "causation".

When GET is followed by an adjective (as in *we got wet*), most often something unpleasant is being reported. In a corpus of spoken English, I found that GET was followed most frequently by "unpleasant" adjectives such as those in [31]. This was a strong tendency, but there was also a small minority of "pleasant" adjectives such as those in [32].

[31] angry, bad, boring, cold, darker, depressed, fat and ugly, jealous, legless, lonely, nasty, nervous, older, paranoid, pissed off, ridiculous, soggy, sore, sticky, violent, worse
[32] better, easier, glad, happy, lucky, warmer

(Cf Biber et al 1999: 481.) A puzzle is therefore how to interpret cases which could be pleasant or unpleasant. For example, depending on circumstances, *getting pregnant* could be a good thing or a bad thing. But there is a norm for collocations with GET plus adjective, and the "unpleasant" connotations may rub off on potentially neutral adjectives. This example merely illustrates the concepts of tendency, instance and norm. In the more complex case of BE- and GET-passives, comparative corpus data are also variable: they show strong, but not absolute, tendencies.

There are two immediate problems for analysis. First, corpus data show that some verbs occur in both constructions:

[33] she thought she was going to be killed
[34] it was mailed just before he got killed
[35] three years later he was arrested in Holland
[36] I didn't get arrested for shop-lifting

Second, passives form a fuzzy set, with central and more peripheral members (Quirk et al 1985: 161; Collins 1996: 45ff; Carter & McCarthy 1999: 7). I will therefore concentrate on central examples which include *by* plus agent, since this guarantees a related active-passive pair, such as

[37] that's where I got hit by a car [A]
[38] that's where a car hit me [M]

BE-passives are at least 30 times as frequent as GET-passives, even in spoken data (Johansson & Oksefjell 1996:69). In addition, BE-passives frequently have agents in *by*-phrases, whereas this is rare with GET-passives. These different frequencies hint at a difference in meaning between the two forms. The much less frequent GET-passive is the marked choice, which suggests that BE-passives tend to be neutral in meaning, whereas GET-passives tend to be chosen for specific communicative reasons.

An analysis I carried out on 10 million words of spoken English shows a strong tendency for GET-passives to be used for talking about unpleasant events. A few examples are:

[39] we nearly got chucked out
[40] customers get embarrassed when talking about money
[41] one child gets hurt
[42] they got kicked out
[43] they got separated from the others
[44] I got walked on by a rather large [...] dog

The BE-passive is certainly also used with "unpleasant" verbs (*accused, banned, barred from, charged with, criticized, dismissed, doomed, forced, murdered, prohibited, upstaged*). However, it is frequently used for neutral and "pleasant" events, as in

[45] they are to be congratulated
[46] the golf tournament was won by Carter

An estimate of how frequently BE- and GET-passives express "unpleasant" events depends on subjective judgements, and figures have to be interpreted generously. However, the differences are large and clear. For a sample of passives (with *by* plus agent) in a 10-million word spoken corpus, my estimate of relative numbers of cases implying unpleasant consequences for the subject-referent is as follows:

GET-passives:	over 60% "unpleasant"; very few "pleasant"
BE-passives:	around 25% "unpleasant"; many "pleasant"

Corpus studies are replicable, and my figures corroborate other studies, which have found that the BE-passive is often neutral in meaning, whereas the GET-passive often indicates that 'something unpleasant is happening' (Francis et al 1996: 58-59; see also Quirk et al 1985: 161; Hübler 1992.) The GET-passive more often expresses emotive or interpersonal meanings, often the speaker's attitude that the event reported is disadvantageous to the subject of the clause, and may also imply that the subject of the clause is responsible for causing an unexpected event (Granger 1983: 196). Collins (1996: 52) and Carter and McCarthy (1999: 49, 50) found nearly 70 and nearly 90 per cent "adversative" uses of GET-passives in a mixed (spoken and written) corpus and a spoken corpus respectively. So, my figure of over 60 per cent for "unpleasant" GET-passives may be rather low, but there is no doubt about the direction of strong regularities which emerge from three independent studies of three independent corpora. The sample concordance lines illustrate the raw data, and readers can judge whether my estimates are reasonable. [NOTE 8.]

CONCORDANCES FOR BE- AND GET-PASSIVES ABOUT HERE

In summary: The patterns are probabilistic. There is no clear boundary between the two passives, but they show strong tendencies to occur in different contexts, neutral versus "unpleasant". There are strong relations between lexico-syntax (BE versus GET), semantics (stative versus dynamic meanings), pragmatics (expression of speaker attitude), and distribution across text-types (formal versus informal). The patterns are not visible in any single instance, but only across many instances in a corpus. This methodological point is clear (Channell 2000: 40). However, the figures are all tendencies, and it is here that the interpretative problems arise.

Now comes a common dilemma. If an analysis is based on only a few examples in their discourse contexts, then it is open to the charge that the data are narrow and unrepresentative. Alternatively, if it is based on a large number of examples, then it is impossible to study the specific context of each one, and the analysis seems superficial. There are different concepts of context, including co-text and inter-text (see above), and norms of usage are often ignored in studies of specific instances. This still leaves us, however, with the problem of how to relate individual instances to the general norm. For example, how do we interpret the rare "pleasant" uses of the GET-passive, such as this one?

[47] I got praised for having a clean plate [A]

Assuming that speakers have internalized a norm of "bad news", can we interpret this as ironic? (See the title and argument in Louw 1993.)

Or do we take this occurrence to be a counter-example? The problem here is that, since there is no absolute difference between the two passives, no single case can be a counter-example to the proposed regularity. There is a corresponding temptation to explain away potential counter-examples on an unrelated set of ad hoc grounds, and a suspicion that appealing to a prototype is a way of ignoring inconvenient data. For example, some cases seem to involve fixed phrases, such as *they got married*. Here a corresponding dynamic form with BE seems only to occur frequently with a time or place adjunct (*they were married on Saturday / at sea*).

Finally, since the patterns are to be found only in large corpora, they are observable only with the help of computer technology. This is no problem in itself: many findings in the natural sciences depend on observations which can be made only with the help of instruments such as microscopes and telescopes. But it leaves unexplained how the unconscious behaviour of individuals can reproduce systematic patterns across a discourse community, and how individual competence relates to social langue.

10. CONCLUDING COMMENTS

W raises important problems of textual interpretation, and his paper forces corpus linguists to be more explicit about the status of the patterns discovered in large corpora. I will conclude by trying to formulate questions which might form the basis for a research programme (compare the lists in Stubbs 1994: 216-18, 1996: 152-53). They would provide empirical evidence, open to testing and falsification, to help answer W's question, namely:

> How can empirical findings about language use be correctly *applied* to problems of textual interpretation?

1. How does frequency of occurrence relate to interpretative significance?

2. Language in use involves both routine and creation. What is the correct statement of this balance?

3. There is often a wide gap between native speaker intuitions and corpus data. What is the relation between I-language and E-language, and how can it be studied?

4. Repeated instances of collocations across a corpus show that meanings are not personal and idiosyncratic, but widely shared. How widespread is such consensus? And what is the relation of these patterns to individual competence?

5. Methods of corpus study, which reveal probabilistic language use, render the concept of a counter-example very problematic. How are instances interpreted when they deviate from the norm?

6. Since words are not distributed in the ways which classic statistical tests assume, it is contentious what tests are appropriate (Dunning 1993, Hogenraad et al 1997). So, what are the appropriate statistical methods for comparing texts with other texts and with corpora?

7. Evaluative and attitudinal meanings are often thought to be due to Gricean inferences, but many pragmatic meanings are conventionally associated with lexico-syntactic structures (Kay 1995, Moon 1998). How much is convention and how much is inference?

Corpus linguistics provides quantities of data which were inconceivable a few years ago, so it is not surprising that these data are now causing problems of interpretation. Corpus linguists think that they have identified a layer of order in these data where none was previously suspected. Widdowson thinks that corpus linguists and critical discourse analysts see more order than truly exists.

11. CODA

Once upon a time, Widdowson was walking in the Scottish Highlands with two colleagues, a critical discourse analyst and a corpus linguist. As they walked, they discussed problems of lexicology, and agreed that it is the phraseology which determines the different senses of *flock* (W pp.17-18) in phrases such as *a flock of sheep* and *holiday-makers flock to Majorca* (not to mention *flock wallpaper*).

Just then a black sheep appeared over a small hill: 'Oh, look!', said the critical discourse analyst, 'Scottish sheep are black.' A few moments later, the rest of a large flock came over the hill: 'Some Scottish sheep are black', said the corpus linguist, rather severely, 'around ten per cent, by the look of it.' Whereupon Widdowson observed: 'There are some sheep in Scotland which appear to be black on at least one side.' And he added: 'But colour is not an interesting property of sheep.' [NOTE 9.]

NOTES

1. On (2): critical discourse analysis is often presented as a contribution to language awareness, and therefore to language education in a broad sense. I have discussed CDA elsewhere (Stubbs 1997a), and agree with many of Widdowson's criticisms. On (3): see Francis and Sinclair (1994), who respond to pedagogical criticisms of corpus-based grammar. And see in particular Sinclair's (1991b) response to Widdowson (1991), which makes clear that there is often much ado about nothing. Sinclair (1991b) 'wholly endorses' (p.491) Widdowson's view of the rights of pedagogy to determine its own affairs, and comments (p.489, 499) that

> 'Corpus linguistics [...] has no direct bearing on the way languages may be presented in a pedagogical context. [...] Corpus linguistics makes no demands on the methodology of language teaching. It is not geared to serving any particular method, and the current software is quite neutral.'

These conciliatory remarks show that some criticism is over a non-issue. Sinclair does then make a fundamental point (p.490) about data: 'many spokespeople in language education are nervous about new evidence, about having to say new and different things about a language'.

2. See also Widdowson (1991), which contrasts 'the possible' and 'the performed' (p.13), accuses corpus linguists of conflating the two (p.14), and implies that corpus linguistics aims to account for 'sentences which happen, incidentally, to have occurred' (p.12). Here, it is the word 'incidentally' which signals a misunderstanding.

3. Widdowson (1996b: 74) points out that 'prototypes cannot be observed', but it does not follow that observational data are irrelevant to identifying them. He provides no evidence for his following statement, that 'conceptual preference [i.e. in prototypes] does not correspond with how frequently these words actually occur'. D'Andrade (1989: 802) reports research on precisely such correlations. (See also W p.19 on prototypes.)

4. The theory of collocational spans is being developed in other empirical work (Sinclair et al 1998, Mason 1999), which studies the exact nature of the span (often asymmetric, of varying lengths, etc).

5. On other aspects of the statistics of text sequence not discussed by either Stubbs (1994/1996) or W, see Hogenraad et al (1997) on auto-correlation: in a text, linguistic features do not occur independently of each other, as is assumed by many standard statistical tests.

6. Stubbs (1997a) is sympathetic to the aims of CDA, but critical of its methods. Stubbs (1997b) discusses sociolinguistic versions of Whorfian arguments, which use quantitative textual data.

7. W discusses only nine individual example sentences. Examples 1 and 2 are identical; 1/2, 3 and 8 are from Stubbs (1996); 4, 5, 6, 7, and 9 seem to be invented.

8. Ikegami (1989) uses corpus data to study the 'prototypical meanings' of constructions with HAVE/GET + object + past participle (as in *HAVE an enquiry carried out*, *GET my glasses changed*). GET is found usually to have a human subject and inanimate object, and to show a high level of agentivity.

9. This second comment of Widdowson's was heard by an anonymous reviewer who happened to be passing at just that moment.

ACKNOWLEDGEMENTS

The data on passives are from a 50-million word sub-corpus of the Bank of English. The analysis of passives was presented at a CUP/BAAL seminar at the Open University in October 1999: I am grateful to Alison Sealey for comments on that occasion. For comments on a previous draft of this paper, I am most grateful to four anonymous reviewers, and to Guy Cook, Gabi Keck, Alison Love, John Sinclair, and Henry Widdowson. Any errors are my own, and Henry Widdowson certainly wishes to reserve his position.

REFERENCES

Atkinson, D. 1999. *Scientific Discourse in Sociohistorical Context*. Mahwah, NJ: Erlbaum.

Baker, M., Francis, G. & Tognini-Bonelli, E. eds. 1993. *Text and Technology*. Amsterdam: Benjamins.

Batstone, R. 1995. Grammar in discourse: attitude and deniability. In G. Cook & B. Seidlhofer eds *Principle and Practice in Applied Linguistics*. Oxford: Oxford University Press.

Bell, A. 1991. *The Language of News Media*. Oxford: Blackwell.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Burrows, J. F. 1987. *Computation into Criticism*. Oxford: Clarendon.

Carter, A. & Sealey, R. 2000. Language, structure and agency: what can realist social theory offer to sociolinguistics? *Journal of Sociolinguistics*, 4, 1: 3-20.

Carter, R. & McCarthy, M. 1999. The English get-passive in spoken discourse. *English Language and Linguistics*, 3, 1: 41-58.

Channell, J. 2000. Corpus-based analysis of evaluative lexis. In S. Hunston & G. Thompson eds *Language in Evaluation*. Oxford: Oxford University Press. 38-55.

Chomsky, N. 1988. *Language and Problems of Knowledge*. Cambridge, MA: MIT Press.

Clear, J. 1996. 'Grammar and nonsense': or syntax and word senses. In J. Svartvik ed. *Words: Proceedings of an International Symposium*. KVHAA Konferenser 36.Stockholm: Almqvist & Wiksell. 213-41.

Collins, P. C. 1996. Get-passives in English. *English World-Wide*, 15, 1: 43-56.

D'Andrade, R. G. 1989. Cultural cognition. In M. I. Posner ed *Foundations of Cognitive Science*. Cambridge, MA: MIT Press. 795-830.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19, 1: 61-74.

Fillmore, C. J. 1992. Corpus linguistics or computer-aided armchair linguistics. In J. Svartvik ed *Directions in Corpus Linguistics*. Berlin: Mouton. 35-60.

Firth, J. R. 1950. Modes of meaning. In *Papers in Linguistics 1934-51*. 1957. London: Oxford University Press. 190-215.

Firth, J. R. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*. Special Volume of the Philological Society. Oxford: Blackwell. 1-32.

Francis, G., Hunston, S. & Manning, E. 1996. *Grammar Patterns 1: Verbs.* London: HarperCollins.

Francis, G. & Sinclair, J. 1994. 'I bet he drinks Carling Black Label': a riposte to Owen on corpus grammar. *Applied Linguistics*, 15, 2: 190-98.

Granger, S. 1983. *The BE + past participle Construction in Spoken English*. Amsterdam: North Holland.

Halliday, M. A. K. 1993. On the language of physical science. In M. A. K. Halliday & J. R. Martin *Writing Science*. London: Falmer.

Halliday, M. A. K & Martin, J. R. 1993. *Writing Science*. London: Falmer.

Hogenraad, R., McKenzie, D. P. & Martindale, C. 1997. The enemy within: autocorrelation bias in content analysis of narratives. *Computers and the Humanities,* 30: 433-39.

Hübler, A. 1992. On the get-passive. In W. G. Busse ed. *Anglistentag 1991 Proceedings*. Tübingen: Niemeyer. 89-101.

Hunston, S. & Francis, G. 2000. *Pattern Grammar*. Amsterdam: Benjamins.

Hymes, D. H. 1972. On communicative competence. In J. Pride & J. Holmes eds *Sociolinguistics*. Harmondsworth: Penguin. 269-93.

Ikegami, Y. 1989. HAVE + object + past participle and GET + object + past participle in the SEU corpus. In U. Fries & M. Heusser eds *Meaning and Beyond*. Tübingen: Narr. 197-213.

Johansson, S. & Oksefjell, S. 1996. Towards a unified account of the syntax and semantcs of GET. In J. Thomas & M. Short *eds Using Corpora for Language Research*. London: Longman. 57-75.

Kay, P. 1995. Construction grammar. In J.-O. Östmann & J. Blommaert eds *Handbook of Pragmatics*. Amsterdam: Benjamins. 171-77.

Kennedy, G. 1992. Preferred ways of putting things with implications for language teaching. In J. Svartvik ed. *Directions in Corpus Linguistics*. Berlin: Mouton. 335-73.

Krishnamurthy, R. 1996. Ethnic, racial and tribal: the language of racism? In R. Caldas-Coulthard & M. Coulthard eds *Texts and Practices: Readings in Critical Discourse Analysis*. London: Routledge. 129-49.

Lee, D. 1992. *Competing Discourses: Perspective and Ideology in Language*. London: Longman.

Levinson, S. C. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

Louw, B. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli eds. *Text and Technology*. Amsterdam: Benjamins. 157-76.

Love, A. M. 1991. Process and product in geology. *English for Specific Purposes,* 10: 89-109.

Mason, O. 1999. Parameters of collocation. In J. M. Kirk ed *Corpora Galore*. Amsterdam: Rodolpi. 267-80.

Moon, R. 1998. *Fixed Expressions and Idioms in English*. Oxford: Clarendon.

Pawley, A. & Syder, F. H. 1983. Two puzzles for linguistic theory. In J. C. Richards & R. W. Schmidt eds *Language and Communication*. London: Longman. 191-226.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Said, E. 1978. *Orientalism*. London: Routledge. [Page refs to Penguin edition, 1991.]

Saussure, F. de 1916. *Cours de Linguistique Générale*. Paris: Payot.

Scott, M. 1997. *WordSmith Tools Manual*. Oxford: Oxford University Press.

Sinclair, J. 1991a. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. 1991b. Shared knowledge. In J. Atlatis ed *Linguistics and Language Pedagogy*. Georgetown: Georgetown University Press. 489-500.

Sinclair, J., Mason, O., Ball, J. & Barnbrook, G. 1998. Language independent statistical software for corpus exploration. *Computers and the Humanities*, 31: 229-255.

Sperber, D. & Wilson, D. 1995. *Relevance: Communication and Cognition*. 2nd ed. Oxford: Blackwell.

Stubbs, M. 1994. Grammar, text and ideology: computer-assisted methods in the linguistics of representation. *Applied Linguistics*, 15, 2: 201-23.

Stubbs, M. 1995. Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of Language*, 2, 1: 23-55.

Stubbs, M. 1996. *Text and Corpus Analysis. Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell.

Stubbs, M. 1997a. Whorf's children: critical comments on critical discourse analysis. In A. Wray & A. Ryan eds *Evolving Models of Language*. Clevedon: Multilingual Matters. 100-16.

Stubbs, M. 1997b. Language and the mediation of experience: linguistic representation and cognitive orientation. In F. Coulmas ed *The Handbook of Sociolinguistics*. Oxford: Blackwell. 358-73.

Teubert, W. 1999. Corpus linguistics: a partisan view. http://solaris3.ids-mannheim.de/ijcl/teubert_cl.html. (Accessed 24 November 1999.)

Tognini-Bonelli, E. 1996. *Corpus Theory and Practice*. Unpublished PhD, University of Birmingham, UK.

Widdowson, H. G. 1991. The description and prescription of language. In J. Atlatis ed *Linguistics and Language Pedagogy*. Georgetown: Georgetown University Press. 11-24.

Widdowson, H. G. 1992. *Practical Stylistics*. Oxford: Oxford University Press.

Widdowson, H. G. 1995a. Discourse analysis: a critical view. *Language and Literature*, 4, 3: 157-72.

Widdowson, H. G. 1995b. Review of Fairclough Discourse and Social Change, *Applied Linguistics*, 16, 4: 510-16.

Widdowson, H. G. 1996a. Discourse and interpretation: conjectures and refutations. *Language and Literature,* 5, 1: 57-69.

Widdowson, H. G 1996b. Linguistics. Oxford: Oxford University Press. Widdowson, H. G. 2000. On the limitations of linguistics applied. *Applied Linguistics*, 21, 1: 3-25.

Concordance. Fifty examples of GET-passives.

The data are from a spoken language sub-corpus of the Bank of English (CobuildDirect). [F0X] etc are speaker identification codes.

```
 1.  Knowing my luck I'll get crushed by a bloody tractor. MX'll be shouting
 2. GY] you're more likely to get hit by a bus walking out this [F0X] Yeah
 3. ould walk out of here and get hit by a car [M01] Right [M0X] I mean you
 4. ant. [M01] That's where I got hit by a car. See that on my knee there.
 5.   you're afraid you might get hit by a golf ball. Right? I think it's ti
 6. ean the same argument get stopped by a policewoman then Frank you you yo
 7.  by the same argument get stopped by a policewoman then Frank [ZF1] you
 8. only person I know who got sacked by a psychotherapist [F01] Mm [F03]
 9. n Bond Street and I got walked on by a rather large and muddy boxer dog
10. my mate yesterday he got attacked by a terrier what was mooching around
11.  were a critique that got coopted by a very different group of people in
12. M01] Well things like getting hit by cars. Falling off the back of a lor
13. apists get fooled and manipulated by clients who are not coming to thera
14. ay I've seen lambs getting killed by dogs. Erm [F02] Killed by dogs [M02
15. dependency that it gets activated by doing a certain amount of drinking
16. F0X] Mm. [F01] Yeah we get funded by er West Midlands Arts and the City
17.     scientists get er get pleased by erm elegant solutions and things of
18.  rather do it erm and I get bored by erm [tc text=pause] because I mean
19. chool gates and that she got done by her mother for just going round the
20. f it. Erm er if you do get struck by Jerusalem recovery is not disastrou
21.   [ZGY] [M01] do you get offended by mother-in-law jokes? [F04] No. No
22.    and erm but it it got reviewed by music critics on the whole erm
23. t I sometimes get a bit irritated by MX who he feels that er now we've g
24. r reason why you would get teased by other people [F04] Well [ZF1] some
25. ep in his car and he got attacked by people with a baseball bat. And er
26. Mm [F03] So she must get accepted by some people more because of that
27.  my Game Boy before it got stolen by some vicious bastard. [M02] Sorry.
28.  she is doing and she gets caught by somebody [F01] Mm [M01] [ZGY]
29. ed their tails if they get caught by something [ZGY] [M01] Uh huh. This
30. name of pub] until we got overrun by students. [F0X] Merchant bankers
31. the rain forest getting destroyed by the acid rain. [F01] And what is ac
32.      erm I think they got stopped by the army or something for just
33. d up I had to go back and got hit by the bouncers. Now what has happened
34.   weren't going to ge get thumped by the er visiting supporters. [M01] M
35.  it don't you. And get frustrated by the fact that you can't do things a
36.  me of people just getting struck by the Holy Spirit. He told me of peop
37. the fascist army and get captured by the partisans who decide to
38.  don't know whether they get paid by the patient or whether he's just
39. when I went to my dad I got dared by the people that the girl that lived
39. old the teacher and they got done by the police for trying to sell us
40. nipulated and getting manipulated by the pop charts and stuff like that
41. . You were just getting barbecued by the power of the Spirit weren't you
43. FO2] Oh [M01] And No you get done by the teacher [F02] Oh [F01] Well hy
44. ool and once again MX got branded by the teachers as lazy and the other
45. e class. [F01] Did you get teased by the teachers in the class? [F02] No
46.  jungle. Only we didn't get eaten by the tiger. [MO1] That's right. [MO2
47.   most they mostly get influenced by their erm parents [F0X] Mm [F05] wh
48. X] Because they're getting backed by their governments to actually do it
49. [F0X] Pervy dirty MX. He got done by t' cops right 'cos [F0X] Yeah. He d
50. ads over there at erm get coached by well ordinary people er do you
```

Concordance. Fifty examples of BE-passives.

```
 1. he Dalkon Shield was manufactured by a company called A H Robbins in the
 2. one point it was going to be done by a Japanese company into a into a
 3.    he and I were both interviewed by a man who wrote a book called The
 4.  since been strongly corroborated by a number of studies [ZGY] which is
 5. ho helped a man who'd been struck by a train near Harrogate and there ar
 6. gineering degrees are now awarded by about forty institutions and some o
 7. space has been rather compromised by an intrusive clutter of parapets
 8.  and he said this kit can be made by any eleven-year-old boy. I'll go I'
 9. generally I mean I was influenced by certain political peoples in my own
10.  range of patterns which are used by doctor and patient to discuss the
11. ls learn effectively being taught by dragons you know. So teaching style
12.  of these assets has been claimed by emergent states and individual repu
13. nd how they might be being shaped by er changes in the N H S. I mean and
14.   airport and being body searched by er the Revolutionary Guards. [M01]
15. ew of the mall which is dominated by erm high-level walkways to left and
16. t that we're now being surrounded by fumes in the j in this little villa
17. here heat is put in are separated by half ocean bases from those places
18.  Yeah [F02] Selby. And I was told by my mother I went with a friend of m
19. hink that was probably stimulated by Nature Conservancy. [MO1] Yes I thi
20. e alpha particle which is stopped by only a few tens of microns [ZF1] of
21. articular risks which are managed by particular companies where I think
22. eenagers are now being questioned by police at Gosport in Hampshire abou
23. they were short and I was invited by Professor MX to come down on a
24. Hollingsworth are being comforted by relatives. This is the update. It's
25. of the deans were firmly squashed by Senate for one reason or another
26. ndustries that are being replaced by some new ones not in any vast
27. Relations [M01] This was produced by that public relations company
28. r er when the police were misused by Thatcher's government. Er do you
29. min which were to be administered by the benevolent city. In such an air
30. us mys er myself. One is employed by the community one is employed by
31. r bit [M01] The men were well led by the Company Commander Lieutenant MX
32.  what is is actually commissioned by the controller and not for us for
33. sions. We were terribly impressed by the courtesy of most of them. Er th
34. hink that the course was affected by the death of FX's husband
35. nced that decisions that are made by the Development Corporation plannin
36. that that confidence is confirmed by the events of nineteen-ninety-six a
37. ar erm weapons that were supplied by the French governments were being
38. elves and what they were supplied by the government so some authorities
39. h erm you know our hands are tied by the National Curriculum [M02] Yeah.
40. NCAR when INCAR was solely funded by the National Science Foundation in
41. s most of that is now been bought by the parish council and there's car
42. ] I dunno whether this is written by the same author but I don't get the
43. ] Yeah [M01] And is that affected by the season? Do you do it at differe
44. of us were at time being detained by the security police and spending ma
45. isms they are are finally humbled by the smallest thing on earth [M05]
46. ones that are going to be cleared by the snow ploughs first and obviousl
47. when this was officially approved by the university and thereafter it wa
48. ] but not being openly advertised by the water company that the office
49.      if you if you are frightened by this person then you have although
50. Friend of Iraq and it is launched by two Kurdish cousins. Their families
```