Unit 12 Objections to corpora: an ongoing debate

12.1 Introduction

A more controversial issue than the issue of representativeness/balance is the question of whether corpus data should be used at all in linguistic analysis, language teaching and language learning. From the 1950s onwards, the corpus-based approach to linguistics was severely criticized, notably by Noam Chomsky (see unit 1.2; see also Aarts 2001; McEnery and Wilson 2001: 5-12). Chomsky's criticism represented an extreme argument against using corpus data. Such a hostile attitude towards corpora has lost credibility in recent years to the extent that the value of corpora is no longer questioned seriously (cf. Nelson 2000). However, amongst those who support the use of corpora there are divergent views on their usefulness for certain purposes. As Murison-Bowie (1996:182) observes, "[t]he strong case suggests that without a corpus (or corpora) there is no meaningful work to be done. The weak case is that there are additional descriptive pedagogic perspectives facilitated by corpus-based work which improve our knowledge of the language and our ability to use it.' The scholars holding the 'strong' view of corpus use include John Sinclair and Michael Stubbs while those who hold the so-called 'weak' view of corpus use include, for example, Henry Widdowson. In this unit we will include three excerpts from published material to give readers an opportunity to understand the different viewpoints and to form their own views. The first excerpt is from a paper by Widdowson, published in *Applied Linguistics* in 2000, where he outlines some reservations he has about the use of corpus data. The second excerpt is a response to Widdowson from Michael Stubbs, published in *Applied Linguistics* in 2001. The final excerpt is cited from de Beaugrande's paper which summarizes the debate between Widdowson and Sinclair over the role of corpora, particularly in language teaching.

12.2 Widdowson (2000)

Widdowson has been mischaracterized as being anti-corpus linguistics. According to Widdowson (personal communication), however, this is not true: 'What I am critical about are the claims that have been made for it, and that's a very different thing, of course.' Widdowson (2000) criticizes some of the claims that have been made in corpus linguistics and CDA (see unit 10.12). The issues about corpus linguistic claims raised in this paper are developed further in Widdowson (2003) while those relating to discourse analysis (especially CDA) are taken up in Widdowson (2004). The excerpt below, cited from Widdowson (2000), focuses on his criticisms of corpus linguistics.

Widdowson, H. 2000. 'The limitations of linguistics applied'. *Applied Linguistics* 21/1: 3-25.

I would argue, then, that linguistics applied is, in effect, misapplied linguistics. And I want now to give some substance to this argument by giving detailed (and critical) consideration to two developments in E-language description that have become extremely influential in our field over the 20 years since this journal was founded. One of these is corpus linguistics: the quantitative analysis of text *en masse*. The other is critical linguistics: the qualitative analysis of particular texts. Each claims to have something quite radical to reveal about language use: corpus analysis about the language that people actually produce, and critical analysis about what they really mean by it. And each also makes claims for the relevance of their analyses to the formulation of problems as experienced in the real world which I believe to be

questionable. In this respect, both are, to my mind, examples of linguistics applied. They warrant close attention because an examination of their analyses and the significance claimed for them seem to me to bring out the issues I have raised in clear relief. Furthermore, the identification of shortcomings (as I see them) at the same time, more positively, points out where applied linguistics might come in.

Corpus linguistics first. There is no doubt that this is an immensely important development in descriptive linguistics. That is not the issue here. The quantitative analysis of text by computer reveals facts about actual language behaviour which are not, or at least not immediately, accessible to intuition. There are frequencies of occurrence of words, and regular patterns of collocational co-occurrence, which users are unaware of, though they must be part of their competence in a procedural sense since they would not otherwise be attested. They are third person observed data ('When do they use the word X?') which are different from the first person data of introspection ('When do I use the word X?'), and the second person data of elicitation ('When do you use the word X?'). Corpus analysis reveals textual facts, fascinating profiles of produced language, and its concordances are always springing surprises. They do indeed reveal a reality about language usage which was hitherto not evident to its users.

But this achievement of corpus analysis at the same time necessarily defines its limitations. For one thing, since what is revealed is contrary to intuition, then it cannot represent the reality of first person awareness. We get third person facts of what people do, but not the facts of what people know, nor what they think they do: they come from the perspective of the observer looking on, not the introspective of the insider. In ethnomethodogical terms, we do not get member categories of description. Furthermore, it can only be one aspect of what they do that is captured by such quantitative analysis. For, obviously enough, the computer can only cope with the material products of what people do when they use language. It can only analyse the textual traces of the processes whereby meaning is achieved: it cannot account for the complex interplay of linguistic and contextual factors whereby discourse is enacted. It cannot produce ethnographic descriptions of language use. In reference to Hymes's components of communicative competence (Hymes 1972), we can say that corpus analysis deals with the textually attested, but not with the encoded possible, nor the contextually appropriate.

To point out these rather obvious limitations is not to undervalue corpus analysis but to define more clearly where its value lies. What it can do is reveal the properties of text, and that is impressive enough. But it is necessarily only a partial account of real language. For there are certain aspects of linguistic reality that it cannot reveal at all. In this respect, the linguistics of the attested is just as partial as the linguistics of the possible.

Problems arise when this partial description is directly applied to determine language prescription for pedagogic use, when claims are made that this provides the only language worth teaching. Now that we know what real language looks like, the argument runs, we expose learners to it and rid our classrooms of contrivance. This follows the common tradition of dependency whereby the language subject is designed in reference to, indeed in deference to, developments in the linguistics discipline. So it was that previously structuralist linguistics defined language content in terms of the formal units of the possible. So it is now that corpus linguistics defines language content in terms of the authentic patterns of the attested. Linguistics applied in both cases. For in both cases, what is not taken into account is the pedagogic perspective, the contextual conditions that have to be met in the classroom for language to be a reality for the learners. Whether you are dealing with the possible or the attested, you still have to make them appropriate for learning. And it is just such conditions that applied linguistics has somehow to take cognizance of.

There are two points (at least) to be made about the direct application of linguistic description of the kind that corpora provide, and both are fairly obvious. The first is that the textual product that is subjected to quantitative analysis is itself a static abstraction. The texts which are collected in a corpus have a reflected reality: they are only real because of the presupposed reality of the discourses of which they are a trace. This is decontexualized language, which is why it is only partially real. If the language is to be realized as use, it has

to be recontextualized. The textual findings of frequencies and co-occurrences have to be contextually reconstituted in the classroom for their reality to be realized, and this reconstitution must obviously be based on very different contextual conditions than those which activated the texts in the first place. The contextual authenticity from which textual features originally derived cannot be ratified by language learners precisely *because* they are learners and do not know (yet) how to do it. It is sometimes assumed to be self-evident that real language is bound to be motivating, but this must depend on whether learners can make it real.

The first point, then, is that however the language is to be contextually abstracted, as units of the possible or the attested, they have to be recontextualized in the classroom so as to make them real for learners. And effective for learning. This is the second point. All language is realized as use in respect to some purpose or other. The purpose of language use in the classroom is to induce learning, and it is appropriate to the extent that it fulfils that purpose. There is a widespread assumption that the classroom is of its nature an unreal place and that this has to be countered by having it replicate the world outside as closely as possible. In the foreign language classroom, this world is taken to mean that of the native speakers of the language concerned. But there seems no good reason why the classroom cannot be a place of created context, like a theatre, where the community of learners live and move and have their being in imagined worlds, purposeful and real for them. To conceive of the classroom in this way is to acknowledge that what is being taught and learned is something designed as a *subject*, not the language as experienced by its native speaker users but something that native speakers cannot experience at all, namely a foreign language. And its foreignness has to be locally accounted for by the devising of appropriate contexts in the classroom which have to activate the process of learning.

This design of the subject is the concern of applied linguistics, whereby descriptive findings are pedagogically treated to make them appropriate as prescription. And the findings of corpus descriptions are potentially highly serviceable to this purpose. It would be a grave mistake to disregard the attested, as it would be to disregard the possible. After all, for many learners at least, the language as realized by its users is the goal to which they aspire and to which they will seek to approximate by the process of gradual authentication. But it would be equally mistaken to suppose that what is textually attested uniquely represents real language and that this reality should define the foreign language subject. A number of people concerned with foreign language pedagogy have expressed reservations along similar lines about the assumption that the findings of corpus linguistics should determine the content of the language subject (Cook 1998; Owen 1993; Widdowson 1991).

It is important to stress that the expression of such reservations does not amount to a denial of the pedagogic potential of corpus description, particularly, perhaps, in the *process* of analysis, in the use of concordancing to develop discovery procedures for learners (see Tribble and Jones 1998; Wichmann et al. 1997). Nor do these reservations betoken a conservative allegiance to outmoded ideas or a stubborn refusal to countenance change, as has sometimes been suggested. Rather it is an effort to refer these descriptive developments to applied linguistic principles by subjecting them to critical appraisal, so as to establish criteria of relevance. It is, of course, important that we should take new modes of description, and their findings, into account in the design of language instruction, but that is very different from accepting them on trust and without question.

To make the point more clearly, let me refer to a particular example. It has been John Sinclair's innovative vision more than anything else that has been the impetus behind developments in the corpus description of English, and he speaks with unique authority as a linguist. He has recently offered a number of precepts for language teachers, the first of which is: *Present real examples only*. These precepts are, as he himself acknowledges, based entirely on descriptive data: 'They are not concerned with psychological or pedagogical approaches to language teaching' (Sinclair 1997: 30). But it seems obvious that if they do not take pedagogic considerations into account, they cannot reasonably be taken as pedagogical precepts. As proposals informed, and so limited, by a linguistic perspective, they may well be worth thinking about, but that is another thing. And to be critically cautious in this way is not

at all to confirm teachers in the belief that they know everything they need to know about the language they teach. But Sinclair thinks it is:

> A few leading figures in applied linguistics (e.g. Widdowson 1992) effectively endorse this complacency by casting doubt on the relevance of corpus findings to the process of teaching and learning languages (Sinclair 1997: 30).

But to cast doubt is to express uncertainty about a claim, not to reject it out of hand. It can only be construed as negative if you assume the self-evident validity of linguistics applied. But from an applied linguistic point of view, casting doubt on the relevance of linguistic description for pedagogic prescription is, I would argue, precisely what we should be about. And this is particularly so in this case. Since, on Sinclair's own admission, the claim for relevance is not informed by pedagogic considerations, it seems only reasonable to entertain some doubt. Indeed, I would argue that the value of such proposed precepts is precisely *because* they provoke a critical response. The alternative is to accept the relevance unilaterally as a self-evident fact, and this means to fashion pedagogic reality to fit the descriptive findings: a clear case of linguistics applied (see also Aston 1995).

I have argued that corpus linguistics provides us with the description of text, not discourse. Although textual findings may well alert us to possible discourse significance and send us back to their contextual source, such significance cannot be read off from the data. The factual data constitute evidence of the textual product: what evidence they might provide of the discourse process is a matter for further enquiry. The same is true, I think, of the other area of description I want to consider. In spite of its name, critical discourse analysis is, I would maintain, also an exercise in text description. And it, too, has a way of assigning discourse significance to textual facts. The supposed area of relevant application is here, however, much broader and of much greater moment. Whereas corpus descriptions have been brought to bear on matters of language pedagogy, critical discourse analysis is concerned with education in a more general sense: it is directed at making people more sociopolitically aware of the way language is used to manipulate them. The purpose is to be applauded. It is hard to think of one which is of greater social significance or more squarely within the scope of applied linguistics. But the question needs to be raised again as to what kind of enquiry this is, and whether here, too, we should take the relevance of its findings on trust, or cast a doubt or two.

12.3 Stubbs (2001b)

As Widdowson's criticisms focus largely on corpus linguistics, Stubbs's response mainly focuses upon the nature of data and methods in corpus linguistics while also addressing the criticism of critical discourse analysis in passing. Stubbs totally bypasses the criticism of the use of corpus data in language teaching. However, readers will have an opportunity to consider this latter criticism in the next section. In the excerpt presented below Stubbs argues that Widdowson's criticism 'is flawed by its misinterpretation of the data, methods and central concepts of corpus linguistics.' Reader can refer to de Beaugrande (2001) for a detailed critical analysis of Widdowson (2000).

Stubbs, M. 2001b. 'Texts, corpora, and problems of interpretation: a response to Widdowson'. *Applied Linguistics* 22/2: 149-172.

THE BACKGROUND ARGUMENTS: DESCRIPTION AND APPLICATIONS

There are two background arguments in Widdowson. 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, Widdowson 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. Widdowson argues that neither perspective, on I-language or E-language, provides the whole truth. Presumably, therefore, they should be combined, though Widdowson makes no proposal as to how this might be done.

Widdowson 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, insofar 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.

THE DATA AND METHODS OF CORPUS LINGUISTICS

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

Possible, attested, and probable

Widdowson (2000: 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 and 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.)

Widdowson'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.

Observational data, introspective data, and mental models

Widdowson 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: 72–3.) This is a valuable and elegant suggestion, but Widdowson does not discuss how these three levels of reality relate to each other, or how such relations could be empirically investigated.

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.

Partiality, point of view, and reality

Widdowson 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 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 (Burrows 1987: 2–3).

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', ibid.: 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 and Sealey 2000). Where I agree with Widdowson is in his insistence that their cognitive influence has yet to be stated clearly.

Interpretation and convention

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

However, one of the deepest problems — which Widdowson 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 and Wilson 1995), and also in a broad tradition of interpretative sociology (to which Widdowson, 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.

Process and product

Widdowson 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 socio-linguistic contexts. Nevertheless, these variable products are due to highly general processes of probability and speaker expectation.

Summary

Widdowson'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.

12.4 Widdowson (1991) vs. Sinclair (1991b): a summary

Stubbs (2001b) does not consider Widdowson's (2000) criticism of the application of corpus data in language teaching simply because he believes that the criticism is 'over a non-issue' and it is 'much ado about nothing' (*ibid*: 170). Yet Widdowson's criticism was not a new one and has been condoned by others (see Seidlhofer 2003: 77-123 for a discussion of the controversy). Indeed the opinions expressed on the usefulness of corpora in language teaching were at least ten years old when Widdowson (2000) was published. As early as 1991, at the Georgetown University Round Table on Languages and Linguistics, there was a heated debate between Widdowson and Sinclair over the use of corpus data in language learning. The paper by de Beaugrande provides a critical review of this debate. The following is an excerpt from his paper.

de Beaugrande, R. Date unknown. 'Large corpora and applied linguistics: H. G. Widdowson versus J. McH. Sinclair'. Accessed on April 16th 2004 at http://beaugrande.bizland.com/WiddowSincS.htm.

In 1991, a controversy arose at the Georgetown University Round Table on Languages and Linguistics during an interchange between Henry Widdowson and John Sinclair. After carefully analysing the two published papers and separately discussing the issues with each of the two linguists, I have concluded that their respective positions are closer together than the controversy might suggest. Widdowson seems to have argued from some positions which are not actually his, and attributed to his opponent some positions which are definitely not Sinclair's.

A predictable crux of the controversy was how corpus evidence might relate to the 'competence' of native speakers on the one hand and to the needs of learners of English as a Foreign Language (hereafter EFL) on the other. As a noted spokesperson for applied linguistics in EFL, Widdowson (1991: 14) felt provoked by Sinclair's typical criticisms, and cited this one: 'we are teaching English in ignorance of a vast amount of basic fact' (Sinclair 1985: 282). To be sure, Sinclair has not blamed the teachers, but the sources they are offered, such as dictionaries, viz:

> Teachers and learners have become used to a diet of manufactured, doctored, lop-sided, unnatural, peculiar, and even bizarre examples through which, in the absence of anything better, traditional dictionaries present the language. It is perhaps the main barrier to real fluency. (1988: 6f)

Nonetheless, Widdowson seemed indignant that 'linguists' who have debarred 'discrimination against languages' should practice 'discrimination against ideas about language'; and that 'linguists have no hesitation in saying that certain ideas held by the uninformed commoner or language teacher are ill-conceived, inadequate, or hopelessly wrong', and in 'rubbishing the theories of colleagues with relish in prescribing their own' (1991: 11). By these tactics, each linguist's 'point of view is sustained by eliminating all others, so that the diversity of experience is reduced in the interests of intellectual security' (1991: 11).

My own detailed studies of the discourse of theoretical linguists in considerable detail (e.g. Beaugrande 1991) confirm Widdowson's remarks. But we should make due allowance for the fact that theoretical linguistics has been largely an enterprise for replacing real language with ideal language existing nowhere except in some 'linguistic theory' (cf. Beaugrande 1997a, 1997b, 1998a, 1998b). In consequence, the major resources for rationally adjudicating theories or models become unavailable, and debaters merely contest that 'my idealisation is better than yours!' At that stage, 'rubbishing the theories of colleagues' and 'eliminating' other 'points of view' become prominent tactics.

The same mode of linguistics would naturally shower 'haughty disapproval, not to say disdain' upon the attempts of 'applied linguists' to 'appropriate' its 'ideas', as Widdowson (1997: 146) has more recently complained (see Beaugrande 1998b for a riposte posted on this website). This posture is not just the ordinary casual 'disdain' of authentic experts for ordinary people. It is the calculated defence of a *sham expertise* that could be severely imperilled by applications, e.g., ones that would quickly debunk Chomsky's (1965: 33) straight-faced denial that 'information regarding situational context' 'plays any role in how language is acquired, once the mechanism' — the 'language acquisition device' — 'is put to work' 'by the child'.

So those earlier polemic tactics ensued from replacing real language with ideal language, whereas the arguments Widdowson was castigating here were being marshalled *against* this very replacement by Sinclair, as they have also been by Pike, Chafe, Firth, Halliday, Hasan, Schegloff, Roy Harris, and many others. Unfortunately, the reinstatement of real language at the rightful centre of modern linguistics cannot be achieved without strenuous 'discrimination against ideas about language' which *really are* 'ill-conceived, inadequate, or hopelessly wrong' but which have been enthroned by linguists whose 'theories' must be sustained by 'rubbishing' the others. And, our own objective is *just the opposite* of 'reducing' the 'diversity of experience' 'in the interests of intellectual security'; we are resolve to *disrupt* the *unearned* 'intellectual security' of linguists, theoretical or applied, who have indeed 'reduced the diversity of experience' of language and discourse and left us with a 'trivial picture' (Halliday 1997: 25).

Widdowson's paper proposed a contrast between the two positions. Whereas the one claims 'objectivity' and 'correctness' in 'descriptions of language', the other adopts 'the relativist or pluralist position on the nature of knowledge':

> The principles or equality and objectivity are comfortable illusions. Descriptions of language are not more or less correct but more or less influential, and therefore prescriptive in effect. They tell us less about truth than about power, about the privilege and prestige accorded to acknowledged authority. […] We cannot any longer be sure of our facts. It is not a very comfortable position to be in. (1991: 11f)

Despite the first person pronouns ('us', 'we'), Widdowson avoided committing himself to this 'pluralist position', but he did imply that Sinclair opposes it by invoking 'basic fact' 'about which teachers were previously ignorant' (Widdowson 1991: 12).

Widdowson then posed the rhetorical question 'what kind of fact is it that comes out of computer analysis of a corpus of text?' (1991: 12). Characteristically, he did not answer it here or anywhere else in the paper by quoting a single 'corpus fact'; at one point, he speculated on the 'relative frequency' of specific words without 'having any evidence immediately to hand' (1991: 17). Instead, he evoked the 'distinction' drawn between 'externalised language' versus 'internalised language' (1991: 12) by none other than Chomsky, the linguist who has memorably taken the most 'relish' in 'rubbishing the theories of colleagues' whilst 'prescribing his own'. Moreover, Chomsky (1991: 89) has 'doubted very much that linguistics has anything to contribute' to 'teaching' (Chomsky 1991: 89), as Widdowson (1990: 9f) has elsewhere acknowledged even whilst rating 'Chomsky's position as consistent with the position I expressed' (but see below). The genuine opposition is still between real language versus ideal language, which, I have asserted, can seriously mislead the language teaching profession.

Widdowson (1991: 12–15) also invoked a further series of oppositions or dichotomies we might do well to deconstruct. These included ‘competence’ versus ‘performance’ (of course); ‘the possible’ versus ‘the performed’ (after Hymes 1972); ‘knowledge’ in ‘the mind’ versus ‘behaviour’ (Chomsky again); and ‘first person’ versus ‘third person perspective’ (Widdowson’ own theme, e.g. 1997: 158f), which should not be misconstrued as referring to the morphology of English Verbs. Sinclair was reproached for conveying the ‘clear implication’ that the corpus is identical with the language, and for excluding the first pole of each opposition whilst allowing only for the second:

> You do not represent language beyond the corpus: the language is represented by the corpus. What is not attested in the data is not English; not real English at any rate. […] what is not part of the corpus is not part of competence. […] What is not performed is just not possible. (Widdowson 1991: 14)

Against this supposed position of ‘the work of Sinclair and his colleagues’, Widdowson quoted Greenbaum (1988: 83) that ‘the major function of the corpus is’ ‘to supply examples that represent language beyond the corpus’. But this position is just as much Sinclair’s, e.g.: ‘language users treat the regular patterns as jumping off points, and create endless variations to suit particular purposes’ (Sinclair 1991: 492). His real position should concur with the notion the collocability and colligability of the lexicogrammar of English are partly realised by the collocations and grammatical colligations of discourse and partially innovated against (Beaugrande 2000).

Sinclair was astounded to be stuck in the straw-man realist position of ‘what is not attested in the data is not real English’ and ‘what is not performed is just not possible’. If he held those positions, he would stop expanding the corpus straightaway because nothing more is ‘possible’ and because any differing data would be ‘not real English’, whereas he has in fact insisted, at times to the dismay of agitated project sponsors, that the corpus must be hugely expanded. He would also have to assume that the sources of his corpus are the linguistic equivalent of the sum total all ‘possible’ sources, whereas he candidly asserts that a much wider selection of spoken data would have already been included but for severe problems of labour and expense.

The evolution of modern linguistics proffers an ironic context for another one of Widdowson’s (1991: 13) polarities: ‘Chomsky’s view is that you go for the possible, Sinclair’s view is that you go for the performed’. By any realistic measure, Chomsky’s programme has always gone for the *impossible*, advocating, with tireless self-confidence, one project after another that never materialise and never could — a ‘grammar’ that is ‘autonomous and independent of meaning’; a solution to ‘the general problem of analysing the process of “understanding” ’ by ‘explaining how kernel sentences are understood’; an account of how human ‘children’ ‘acquire language’ by ‘inventing a generative grammar that defines well-formedness and assigns interpretations to sentences even though linguistic data’ are ‘deficient’ (1957: 17, 92; 1965: 201); and more others than I have room to list here (for a thorough analysis of Chomskyan discourse, see now Beaugrande 1998b).

Here we can look to Hjelmslev (1969 [1943]: 17) for the most striking formulation, this one concerning the ‘possible’: ‘the linguistic theoretician must’ ‘foresee all conceivable possibilities’, including ‘texts and languages that have not appeared in practice’ and ‘some of which will probably never be realised’. Easy enough to say once you decide (as we saw Hjelmslev do) that ‘linguistic theory cannot be verified (confirmed or invalidated) by reference to any existing texts and languages’.

Chomsky (1965: 25, 27) fulfilled Hjelmslev’s vision in the most facile manner when he simply installed, by fiat, just such a ‘theory’ in the ‘language acquisition device’ of the human child: ‘as a precondition for language learning’ the child ‘must possess a linguistic theory that specifies the form of the grammar of a possible human language’ plus ‘a strategy for selecting a grammar’ by ‘determining which of the humanly possible languages is that of the community’. This is definitely not the position of Widdowson, who has firmly rejected the

concept of ‘internalisation’ by means of a ‘universal Chomskyan language acquisition device’ (1990: 19).

The conception of the ‘possible’ is too abstract to be very useful for language pedagogy anyway. Learners of English as a non-native language produce many utterances which may not seem possible to the teacher’s intuition, but, as I have noted, we are currently finding new motives for doubting the reliability of intuition. Far more relevant is what is or is not both ‘possible’ and ‘performed’ at the learners’ *current stage of skills and knowledge*, since that is all we can realistically hope to build upon. There, we can productively orient our approach toward large corpora of *learners’ English*, such as have been collected by Sylviane Granger at the University of Louvain (cf. Granger 1996) and by John Milton at the Hong Kong University of Science and Technology (cf. Milton and Freeman 1996). Such data can also systematically alert teachers and learners to typical problems such as language interference.

Another of Widdowson’s polarities we might deconstruct is the one between ‘knowledge’ in ‘the mind’ versus ‘behaviour’, the latter term perhaps reminding language teachers of behaviourist pedagogy and Skinnerean behaviourism. But linking a large corpus with behaviour and behaviourist methods would be flawed for at least two reasons. The more obvious reason is that the behaviourist ‘audio-lingual’ method with its pattern drills and prefabricated dialogues was based on mechanical language patterns more than on authentic data; it equated language with behaviour in order to reduce language, whose relative complexity it could not grasp, to behaviour, whose relative simplicity seemed ideal for ‘conditioning’, ‘reinforcement’ and so on; and the method was backed up by heavy behaviourist commitments with in general pedagogy and by the prestige and authority of American military language institutes, where ‘drills’ are literally the ‘order of the day’. Nor does Sinclair advocate a teaching method whereby learners parrot back corpus data; on the contrary, he has expressly counselled against ‘heaping raw texts into the classroom, which is becoming quite fashionable’, and in favour of having ‘the patterns of language to be taught undergo pedagogic processing’ (1996).

The more subtle reason is that corpus data are not equivalent to ‘behaviour’ in the ‘externalised’ sense which Widdowson’s polarities imply and which is often encountered in discussions of pedagogy, e.g., when a ‘syllabus’ ‘identifies’ ‘behavioural skills’ (Sinclair 1988: 175). Instead, they are *discourse*, and the distinction is crucial. External behaviour consists of observable corporeal enactments, of which the classic examples in behaviourist research were running mazes, pulling levers, and pressing keys. Discourse is behaviour in that externalised sense only as an array of articulatory and acoustic operations, or, for written language, of inscriptions and visual recognitions; and no one has for a long time — certainly not Sinclair — proposed to describe language in those terms, nor does a corpus represent language that way. When discourse realises lexical collocability and grammatical colligability by means of collocations and colligations, the ‘performed’ continually re-specifies and adjusts the contours of the ‘possible’. In parallel, ‘knowledge’ in ‘the mind’ decides the *significance* of the ‘behaviour’. Sinclair’s true position is that these operations are far more delicate and specific than we can determine without extensive corpus data. Moreover, analysing corpus data is less equivalent to *observing behaviour* than to *participating in discourse*.

12.5 Unit summary and looking ahead

This unit presented three excerpts from published material in order to explore the pros and cons of the corpus-based approach. These characterize well the so-called strong and weak cases for the use of corpus data. Between these two poles are many milder (positive or negative) reactions to corpus data (see Nelson 2000 section 5.3.3 for an overview of these reactions). Nevertheless, the discussion in this unit clearly shows that while some reservations remain about the use of corpus data, corpora have generally been accepted as valuable linguistic resource. Readers are reminded that the corpus-based approach and the intuition-based approach are not conflicting but complementary (see unit 1.5). We have already noted that the corpus-based approach

is not all-powerful (cf. unit 10.15). The usefulness of a corpus is typically dependent on the research question researchers intend to address using the corpus. Also, corpora do not necessarily provide explanations for what we see. This remains the task of the human analyst, drawing upon a wide range of resources, and methodologies. Nevertheless, corpora are undoubtedly valuable resources in linguistic analysis and language teaching, as has been shown in unit 10 and will be demonstrated in the second part (units 13-16) of Section B.