A two-way exchange between syntax and corpora

In his contribution, Geoffrey Sampson, Professor Emeritus at the University of Sussex (United Kingdom), highlights the relationship between Corpus Linguistics and Syntax. He shows how this bond has a two-way nature. In his view, the use of corpora in language research allows one to better understand syntactic issues and the development of language complexity. However, the other way is also true in Sampson's view since he believes the focus on syntax is one of the major factors contributing to the growth of interest in Corpus Linguistics. From a more general perspective, Sampson argues in favor of linguistics remaining a creative activity which develops in unexpected ways. As for the prospects of Corpus Linguistics, he predicts its death – not of this approach itself, but of the term. He believes the label 'corpus linguistics' will disappear when corpora become just another resource available to linguists.

1. **Where do you place the roots of Corpus Linguistics? And to what do you attribute the growth of interest in the area?**
2. **Is Corpus Linguistics a science or a methodology? Where would you situate Corpus Linguistics in the scientific or methodological panorama?**

I must take these questions together, because answering either one involves discussing the other.

The first thing that needs to be said about these and the rest of this series of questions (I shall be surprised if I am the only contributor who makes essentially the same point) is that it is misleading to think of "Corpus Linguistics" as a branch of linguistics, alongside sociolinguistics or historical linguistics. Corpus linguists are just people who study language and languages in an empirical, scientific manner, using whatever sources of empirical data are available; at the present time it happens that, for many aspects of language, the most useful data sources are often electronic corpora. I work a lot with corpora, but I think of myself as a linguist, not a "corpus linguist". If some aspect of language is better studied using other tools, I will use those.

The reason why corpora have become more significant in linguistics than they used to be include: (i) the availability of computers; (ii) change of emphasis from phonology to syntax; and (iii) the bankruptcy of intuition-based techniques. I discuss these points in turn:

Availability of computers

It is hard to do much with a corpus unless it is in electronic form and you have access to a computer to process and search it. The Brown Corpus, the first electronic corpus, was published in 1964, which as it happens was close to the time when I began learning to work with computers – but that was very unusual then for someone with a humanities background. Everyone had heard of computers, but most academics knew little about them and had certainly never seen one. I remember the air of imperfectly-concealed condescension with which engineers and mathematicians greeted the idea that some of us arts types wanted to play with their machines. When we managed to do so, the low-level programming languages of those days and the batch-processing approach of 1960s computing environments meant that, although one could use computers to find out things about language which would be hard to discover any other way, the process was horribly slow and cumbersome relative to what is possible and easy now.

It was not until some time in the 1980s that computers began to become routinely available to linguists. Even that is quite a while ago now; but when a complex new technology does become convenient and widely available, it inevitably takes time for a profession to adjust to its possibilities. Corpus-based techniques have taken decades to catch on in linguistics, but I am not sure that one could have expected the process to occur faster.

Change of emphasis within the discipline

Until some point in the 1960s, the intellectual "centre of gravity" of linguistics lay in phonology, which deals mainly with finite systems of a few dozen phonemes that combine in a limited number of ways. Corpora do not offer much to the phonologist. One can survey the possibilities adequately using traditional techniques. Only with the rise of generative linguistics did the "weight" of the discipline shift to syntax, which deals with large numbers of elements combining in effectively infinitely many ways. That meant that one needed to study very large samples to have a chance of encountering a representative range of possibilities, so corpus compilation became the way forward.

Bankruptcy of intuition-based techniques

Ironically, while the generative movement shifted linguists' attention to an aspect of language – syntax – which is difficult to study empirically without the use of corpora, the unempirical style of research advocated by the generativists led very many linguists to ignore the virtues of corpora for a long time after they started becoming available. No-one in the modern world would suggest that, say, meteorologists or marine biologists should decide what their basic data were without looking at evidence: it is too obvious that the weather, and marine organisms, are things independent of us and that we can find out about them only by looking. Language is not in the same sense independent of human cognition, so it may at first have been reasonable for the Chomskyans to believe that a linguist can decide what is in and what is not in his language by introspection, without external observation. And, as well as arguing that grammar-writing *can* be based on introspection, they cited the "absence of negative evidence" (that is, we don't hear starred sentences) in order to argue that grammar-writing *cannot* successfully be based on observation.

For a short while these ideas may have been reasonable, but it soon turned out that eliminating the dependence of science on observation is just as bad an idea in linguistics as in physical sciences. This was clear at least from the time when William Labov (1975) demonstrated that speakers simply do not know how they speak, and that generative linguists ascribe an authority to their own judgements which they manifestly do not possess. The argument from absence of negative evidence represented a misunderstanding of how empirical science works (Sampson 1975); if it were a good argument, no physical science would be possible (Sampson 2005: 89–91).

By now there are many cases where core elements of non-empirical linguists' theories rest on intuitive beliefs that are wildly at variance with reality. One of Noam Chomsky's leading arguments for innate knowledge of language (see e.g. Chomsky 1980: 40) is the claim that, without innate knowledge, children could not succeed in mastering the English rule for forming questions, because structures that are allegedly crucial for determining the correct rule are so rare that one can live one's life without ever hearing an example. Chomsky seems to have based that statement on guesswork (or "intuition", if one wants to use the more dignified term). Although I do not believe that one needs to hear these particular structures to get the question rule right, I used the demographically-sampled speech section of the British National Corpus to check how rare the structures are in real life. It turned out that one can expect to hear thousands of relevant examples in a lifetime's exposure to casual chat (Sampson 2005: 81). This is not an isolated case of mismatch between generative linguists' intuitions and empirical

reality (though it is perhaps the most egregious case, in view of the frequency with which the generative literature has relied on this baseless assertion – cf. Pullum and Scholz 2002: 39–40).

Even in face of absurdities like this, quite a few linguists do continue to cling to the idea that grammatical research can progress independently of empirical evidence. But by now they are starting to resemble upper-middle-class Edwardian ladies who cannot conceive of cooking or cleaning with their own hands. Fiddling about with scripts for searching text files or with tape recordings of spontaneous speech looks like servants' work to some of the more precious inhabitants of linguistics departments. But the reality of many areas of present-day linguistics is that, if one wants to make progress rather than just go through the motions, that is the kind of work that has to be done; and I think this is now obvious to many younger linguists. So it is no surprise that corpus work has been coming to the fore.

The remaining point in Questions 1 and 2 concerns the "roots" of corpus linguistics. Diana McCarthy and I surveyed the historical origins of corpus work briefly in our *Corpus Linguistics* anthology (Sampson and McCarthy 2004: 1–4). One might argue that Dr Johnson's dictionary was based in part on a "corpus" of literary quotations, and the work of Wilhelm Kaeding (1898) seems to have been a clear early case of corpus linguistics in the modern sense. But these are matters of fact and of definition (what counts as a "corpus"?), rather than of intellectual controversy; there is little to be gained from contributors repeatedly rehearsing the history at length.

3. How representative can a corpus be?

Representativeness seems to have become something of a bugbear for corpus researchers, but I am not quite sure why it is felt to be a worry. Any corpus is a sample of language use, and naturally one wants it to be an unbiased "fair sample". Statisticians who discuss sampling talk in terms of drawing a sample from a "population" – the (perhaps infinitely) numerous set of entities for which the finite sample is intended to stand proxy. If there is a worry about corpus representativeness, perhaps the problem is less about sampling techniques than about deciding what "population" is to be sampled. Thus, for written language ought we to think in terms of acts of writing, or acts of reading (some pieces of written language are read very many times, others only once)? Or perhaps the problem arises because of tensions between groups who want to use language corpora for different purposes and have not fully recognized that the same kind of sample will not suit all purposes equally. The written-language section of the British National Corpus includes quite a lot of literary writing, sometimes decades old. For a sociolinguist interested in what written usage the average Briton encounters, this

might be inappropriate; for the dictionary publishers who were among the leading sponsors of the BNC project, it may be very desirable to give extra weight to writing that is recognized as more authoritative than, say, hastily-composed office memos. This would be a case of conflicting interests; I wonder whether "representativeness" is invoked in order to suggest that such conflicts have scientifically correct solutions.

To me it is hard to get worked up about this issue, because (at least with respect to English grammar, the aspect of language that I have chiefly been involved with) such evidence as I have examined suggests to me that any differences between genres of English are trivial relative to what they have in common (Sampson 2001: Chapter 3). There is one English language, not a set of Englishes. Clearly we should avoid obvious bias in the way we sample the language, when we can easily do so, but I am sceptical about whether our findings will be much affected by how far we go to achieve perfect representativeness.

4. How far should an analyst rely on intuition?

I have discussed intuition to some extent in an earlier answer. The standard line, according to the hypothetico-deductive scientific method, is that the scientist uses intuition to generate plausible hypotheses – hypotheses will not emerge mechanically from any amount of accumulated data – and then uses empirical evidence to corroborate or refute the hypotheses. This is as applicable to linguistics, I believe, as to other fields, and in linguistics the empirical evidence often comes from corpora.

5. What kind of questions should an analyst think of?

This one really is unanswerable! Linguistics is a science, and science is a creative affair – a scientist who hopes to be told what kind of questions to ask (what hypotheses to formulate, in the jargon) is unlikely to produce much of value.

Admittedly that might not be true of present-day "Big Science": 21st-century genetics, for instance, seems to involve armies of researchers uncovering and assembling numerous small pieces of new knowledge in response to strategic research guidelines which perhaps can be laid down successfully well in advance. Whoever formulated the guidelines needed to be creative, but the individual researchers possibly do not. However, linguistics, realistically, will never be like that (and probably should not be like that even if it could be). Linguistics will always be "craft science" rather than production-line science, organizationally more like seventeenth-century physics than 21st-century Big Science. That means that it is heavily dependent on individuals with original minds spotting novel questions whose answers might move our understanding forward. One cannot lay down

long-term research strategies, because tomorrow's questions grow in an unpredictable fashion out of today's answers.

Now that the management of universities is increasingly shifting out of the hands of practising academics into those of professional managers, these points are beginning to be lost sight of. In my experience the managerial types in suits would like university research to move into predictable, production-line mode, and they have little understanding of (or patience with) the idea that for many subjects it just cannot be like that. Younger academics, whose memory does not stretch back to a time when university governance was in a healthier state, are sometimes browbeaten into accepting that the managerial perspective must be correct. But, for linguistics, the "production line" research model could only be a system for raising and spending funds in an orderly manner and providing researchers with a career structure. If it generated any significant advances in our understanding of language and languages, these would surely emerge more or less accidentally, out of the tea-breaks or things done after the end of the shift, as it were, rather than rolling systematically off the end of the production line.

Whether as a consequence of managerialism or for other reasons, linguists who work with corpora do often seem to misunderstand the essentially creative aspect of the discipline. One symptom of this is the way that groups who publish new corpus resources are routinely expected nowadays to complement the data files with software for manipulating them. When I began to work with the British National Corpus and subscribed to its online forum, I was surprised (and quite disappointed) to find that it was full of messages about how to implement the software accompanying the BNC (called Sara, if I remember correctly), while there was hardly anything about people using the BNC to explore the nature of the English language in novel ways. I have even had people complain to me that the corpus resources I have made available to the public are only half-finished, because I provide no software to go with them – though I do provide documentation which defines their file structures very precisely.

Personally, when I get hold of new corpus resources, I use the data files and discard or ignore any software that comes with them. However good the software might be, it will be designed to allow users to answer some fixed range of questions which the designer anticipates that people will want to ask. The chances that this range will cover the questions I find myself wanting to put to the data are not good enough to make it worth learning to use the software. Clearly that cannot be an absolute rule: when the recordings underlying the spoken section of the BNC are digitized by the "Mining a Year of Speech" project which John Coleman is leading at Oxford, I shall have to use that project's software to explore the material – it would be folly to try to analyse acoustic signals independently. But most electronic corpora, from Brown and LOB to the existing BNC, comprise

straightforward text files, so that it is easy to write one's own scripts to analyse them in whatever way one wants. If a linguist is not willing to learn enough Perl to write simple analytic routines, then I'm sorry, but he or she is in the wrong job.

Now that corpus development has become a widespread activity within the discipline, one is hearing complaints by sceptics that for all the effort going into corpus-building, there does not seem to be a commensurate volume of new knowledge and insights emerging from corpora. I have sympathy with this complaint. At the present juncture I have a sense that there are a number of linguists around the world who like the idea of getting funding to develop a corpus of their language or their favourite genre of language use, but who do not really look beyond the busy-work of getting the corpus compiled; they perhaps hope vaguely that when their corpus exists, valuable knowledge will emerge from it almost automatically. That won't happen. A corpus is only a tool, and there is little point in equipping oneself with an expensive tool unless one has plans for using it.

6. What are the strengths and weaknesses of corpus analysis?

A "corpus" just means a collection of samples of language usage recorded in some manner or other. If one is tempted to say that language corpora are unsuitable for certain kinds of linguistic research, one must be careful that the appearance of unsuitability does not merely reflect unduly narrow assumptions about the nature of corpora. For instance, traditional corpora of transcribed speech might not be adequate for studying child language development, even if the speech is that of children, because one cannot see what the child is doing or what is going on around him as he speaks. But a collection which videotaped the scenes as well as recording the sound would still be a "corpus", though one very different from the classic language corpora.

Nevertheless, it is true that corpora are more useful for some areas of linguistic research than others. When we are dealing with small finite systems (the phoneme systems already mentioned being the obvious example), corpora tend not to be needed; we can often get on fine without them (though there are aspects of phonology, notably intonation systems, where corpus work will often be valuable or essential). At the other "end" of linguistics, it seems to me that corpora have limited relevance (though some relevance) to the study of semantics. But that is not because the semantics of a language is studied using other sources of empirical evidence which do not fit the definition of "corpus". It is because to a large extent the study of semantics is not an empirical scientific discipline at all, but something more like a branch of philosophy (cf. Sampson 2001: Chapter 11). Subjects which can be studied scientifically ought to be studied that way, but we must recognize that science has limits.

7. What is the future of Corpus Linguistics?

As suggested in my previous answer, for the immediate future the priority needs to be (and I hope will be) a shift of emphasis, away from creating yet more corpora, towards extracting worthwhile knowledge from those we already have. By now we have lots. Of course I understand that if you are from a country whose national language has no corpus at all yet, building its Brown/LOB equivalent will be a high priority. But searching out and seeking to fill increasingly narrow "gaps in the market" strikes me as a questionable use of linguists' time. The world does not truly need, say, a corpus of informal conversation between legal professionals (an example which I hope is hypothetical – no offence to anyone is intended). If the legal profession needs such a resource, let them take the initiative towards compiling it; they presumably will know how they want to use it. At present, the existing array of corpora are underexploited, so our profession ought to be putting effort into formulating novel questions to put to them.

Looking a little further ahead, in a sense I believe that corpus linguistics as such has not got a future. I began by saying that "Corpus Linguistics" is not a special branch of linguistics. I would hope that its future is simply to fade away as a concept, because all concerned will take corpora for granted as one important set of tools in any linguist's toolbox. Some linguists will work with corpora most of the time, others more sporadically, and no doubt some will specialize in areas where corpora have little or no relevance. But it seems to me that it will be quite a failure if in forty years' time the phrase "corpus linguistics" continues to be an established collocation.

8. What issues does one have to face when developing treebanks?

If we want to use a corpus to find out about aspects of a language other than vocabulary, we will probably need it to be equipped with annotation making explicit the grammatical structures into which the words are organized. Almost from the beginning of electronic corpus compilation it was usual to add part-of-speech tags to the words, and for a long time now many corpus developers have been adding information about phrase and clause structure – turning raw corpora into "treebanks".

The biggest problem here lies in taxonomy. What range of syntactic structures does a language possess, and where are the boundaries to be drawn between different categories of constituent? Linguists who became used to the aprioristic syntactic theorizing of the 1960s and 1970s learned a few standard categories – noun phrase, adjective phrase, complement clause, relative clause, and so on; but, as soon as one encounters real-life language samples (even if these are drawn from

edited, published writing, let alone from casual speech) one is rapidly at a loss to know how to apply the familiar categories, or to decide what further categories should be postulated. What labelled bracketing should we assign to a postal address? In the sequence *we kept adding to our ritual without daring to abandon any part of it*, is *without* functioning as a preposition introducing a separate constituent headed by *daring*, or is *without* a subordinating conjunction acting as the first word of a non-finite clause, parallel to, say, *while seeking to* …? In my experience, one begins treebank compilation imagining that after a few debatable issues like these are cleared out of the way, the rest of the work will be fairly plain sailing – but it does not take long to discover that the debatable issues are almost more numerous than the straightforward cases, and new debatable issues never stop cropping up.

The point was demonstrated experimentally at a workshop at the 1991 ACL annual conference. Computational linguists from nine institutions were given a set of English sentences and asked to indicate what bracket-structure their respective groups would assign to them; and the analyses were compared. They were not asked to label the brackets; it is easy to imagine that different groups might use different nomenclature for grammatical categories even if they meant essentially the same thing. But one might have expected that at least the placing of brackets would agree fairly well. Yet, although the sentences were not notably "messy", agreement was strikingly poor. In the following sentence (from a *New York Times* article included in the Brown Corpus):

> One of those capital-gains ventures, in fact, has saddled him with Gore Court.

the *only* constituents identified as such by all nine participants were the name *Gore Court*, and the prepositional phrase *with Gore Court*.

In this situation it seems inescapable that if we want to get anywhere with building meaningful treebanks and generating findings that can meaningfully be shared between research groups, a high priority must be to define analytic schemes that will not just specify a comprehensive range of categories but will offer detailed, rigorous guidelines specifying how they are to be applied to as many debatable cases as possible. Our situation is akin to that confronting Carl Linnaeus when he developed the first standard system for naming biological species. Without it, there was just no way for botanists in different places to know whether or not they were discussing the same plant.

When, with colleagues at Lancaster University, I began developing what I believe may have been the first-ever treebank in the early 1980s, rigour and comprehensiveness in the analytic scheme seemed to me a more important goal than size of the treebank. Although scheme and treebank grew in parallel, I think the

wordage of the scheme definition was always substantially in excess of the wordage of the analysed samples comprising the treebank. But, as treebank development has become an international industry, others involved in it do not always seem to have seen things the same way. My impression is that commonly it is seen as much more important to produce the largest-possible treebank than to adopt rigorous definitions of the analytic categories.

Academics are at the mercy of research sponsors, of course, and in dealing with funding agencies it is undoubtedly easier to "sell" an enormous treebank than a tightly-defined treebank. Yet, without tight definition, the larger the treebank the more likely it is that its annotations will not reliably be counting apples with apples and oranges with oranges. Research sponsors may not initially appreciate this problem, but it is our role to educate them.

There is of course a long history of downplaying the importance of taxonomy in linguistics. Generative linguists have in the past expressed hostility to taxonomy. Consider for instance Jerrold Katz's comments (1971: 31ff.) on linguistics as "library science", as he put it, or the negative connotations of Chomsky's use (1964: 11) of the term "taxonomic model". And now that the generativists have moved on from NP and VP to "Spec C", "TP", and their other latter-day syntactic symbols, they seem to have shifted, if anything, even further away from the nitty-gritty issues of "Where exactly does this unusual-looking constituent begin and end, and how do we classify it?", which constantly face anyone who tries to turn a real-life corpus into a treebank.

No corpus linguists, I think, are actually hostile to the taxonomic enterprise. The point I am trying to make in this section was made with more eloquence than I can muster by Jane Edwards at the corpus-linguistics Nobel Symposium (Edwards 1992: 139):

> The single most important property of any data base for purposes of computer-assisted research is that *similar instances be encoded in predictably similar ways*.

But this is a principle which I feel the community of corpus analysts in general has not yet taken fully to heart. Defining detailed, comprehensive analytic guidelines is an unglamorous, indeed downright tedious activity, but it merits a larger share of corpus linguists' efforts than it has been receiving.

9. In what way(s) can Corpus Linguistics enhance our understanding of syntax? And how is it reflected in grammar books?

As already suggested, syntax is to my mind the aspect of language where corpus-based research is supremely useful. It can answer questions that could scarcely be addressed any other way, ranging from highly specific queries such as whether

some individual construction remains current, or what features in the environment favour or disfavour its use, to very general issues about the nature of human language behaviour.

One of these general issues which corpus work has led me to see with new eyes concerns the concept of "ungrammaticality".

For half a century now, most theoretical linguists have understood the grammar of a natural language on the model of the artificial "languages" of mathematical logic, such as the propositional calculus, where the concept *well-formed formula* is central to the system. Rules generate an (infinitely numerous) class of symbol-sequences that count as meaningful formulae of the calculus; other sequences of the same symbols are meaningless jumbles. Linguists, similarly, have identified a language such as English with an (infinitely numerous) class of grammatical English sentences. They have seen the task of grammar-writing as being in large part to devise rules to distinguish between the grammatical sentences and the "starred strings" or "word-salad".

Linguists have always recognized that the rules of grammaticality for a natural language must be massively more complex than those of artificial formal languages. And they have nuanced the picture in further ways. Some linguists suggest, for instance, that rules of natural-language grammar should be supplemented with probabilities, or with information about social variables, so that rather than merely defining a two-way grammatical/ungrammatical classification of strings of words, the grammar might characterize a sentence as "grammatical but unusual", or "used by men more than women". But the idea that below this detail there is a fundamental distinction between grammatical and ungrammatical, whatever type of rules may be needed to formalize that distinction, has scarcely been challenged. For many years I took it for granted myself.

There were always linguists who questioned the orthodoxy. Fred Householder (1973: 371) pointed out that it is remarkably difficult to construct a sequence of English words for which one cannot imagine any use whatever. Studying the statistical distribution of constructions in English-language corpora eventually convinced me that the concept of "ungrammaticality" is fundamentally mistaken. I no longer believe that any two-way classification of that sort can be imposed on word-sequences; the analogy with logical calculi is severely misleading.

Clearly, any language has some grammatical constructions which are very familiar and heavily used, and others which are less standard but will be used on occasion – but the evidence I have seen suggests that this is a cline with no particular termination. In a "target article" in the "Grammar without grammaticality" special issue of the journal *Corpus Linguistics and Linguistic Theory* (Sampson 2007) I discussed this evidence, and likened the situation as I now see it to the pattern of tracks in open savannah country inhabited by a population which has

not developed formal systems of land law, rights of way, and so forth. There will be some wide, heavily-used roadways, other lesser tracks, and so on down to scarcely-visible marks in the grass where one or two pairs of feet have passed. But it will not make sense to ask "Is there a track from point X to point Y?" – in the imaginary scenario I postulated, if X to Y does not coincide with a heavily-used route the answer would have to be something like "I don't remember seeing anyone walking just that way, but if you want to, go ahead".

Similarly in the case of language, it makes sense to ask "Can one say *The farmer killed the duckling* in English?", and the answer will be "Yes, subject–verb–object is one of the central sentence-patterns of the language", but it does not really make sense to ask "Is XYZ ungrammatical?" – if XYZ is a peculiar string of words, the only reasonable answer would be something like "Well, what do you mean by it? – of course if that is how you want to use it, nothing stands in the way". The situation is quite different from the case of the propositional calculus, where permuting the symbols of a well-formed formula gives a sequence that is just meaningless and useless, full stop.

Many commentators on my target article disagreed with me; but much of the disagreement read more as if the commentators could not believe that I was serious about holding such an unorthodox position, than as if they understood what I was saying and believed it was mistaken for identifiable reasons. The ungrammaticality concept appears to have such a hold over present-day linguistics that people find it difficult to entertain the possibility that it is a mistake.

Yet it is a fairly recent concept. The asterisk notation for ungrammaticality was never used, so far as I know, before the rise of generative linguistics. (I believe it was adapted from historical linguists' use of the asterisk to indicate that a reconstructed form is not actually attested – a quite different concept.) The "pedagogical" or "descriptive" grammar books that have been published down the centuries, before theoretical linguistics existed, listed constructions that do occur in a language but it seems to me that they did not express (or imply) any complementary concept of impossible constructions or word-sequences. If they mentioned that some form of wording was to be avoided, that was because people *do* often use it but it is socially deprecated.

In this respect it seems to me that descriptive grammars of languages, which theoretical linguists have sometimes seen as anecdotal or intellectually lightweight relative to their own attempts to formalize grammar rules, are more faithful to the reality of human language than a formal grammar can be. Someone who made a map of tracks in the savannah would include the broadest paths and some of the lesser ones, but would have to choose an arbitrary cut-off point below which paths were too narrow and temporary to mark on the map. Descriptive grammar-books do something very like that for natural languages: they list the heavily

used constructions and some of the less heavily-used ones, and it is an arbitrary decision where to stop and treat more unusual forms of wording as too occasional or specialized to mention.

Without corpus experience, I personally would probably never have come to see language this way. Perhaps it is no coincidence that Geoffrey Leech, one of the co-authors of the best-established descriptive grammar of English (Quirk et al. 1985), was also the pioneer of corpus linguistics on our side of the Atlantic. My current "take" on the ungrammaticality concept may itself be misguided, of course – but I cannot imagine what category of evidence other than corpus evidence could be used to construct a serious argument against it.

10. What do corpus-based studies tell us about the development of language complexity? And how have/should they impact language teaching?

If one believes, as Noam Chomsky and Steven Pinker do, that the overall architecture of human language is laid down in our genes, then development of language complexity is scarcely an issue. One would expect all human languages to be similar in structure and hence similar in complexity, and an individual's idiolect would not be expected to develop much in complexity after he or she has passed the "critical period" when the innate Language Acquisition Device is biologically programmed to switch off. But we know more about genetics now than we did when Chomsky was developing his ideas about innate knowledge of language, or even than we did when Pinker wrote *The Language Instinct* (Pinker 1994), and it has become harder to see how their picture of language acquisition could possibly be correct. (Chater et al. 2009 have produced a formal argument that it cannot be correct, though some have rejected that argument.)

Whether language structure *could* be genetically encoded or not, I find Chomsky's and Pinker's arguments that it *is* so quite empty (Sampson 2005); and others (notably Evans and Levinson 2009) are independently drawing attention to the fact that patterns of diversity among languages seem incompatible with the Chomsky/Pinker picture. The reasonable conclusion at this point is surely that languages are cultural constructs, constrained only in minor respects by biology. In that case, one would expect to find differences in complexity among languages, growth in complexity over time, and so forth, as one finds in other areas of human culture.

One way in which I have brought corpus data into relationship with this idea was by looking at correlations between syntactic complexity and speakers' demographic characteristics in a subset of the BNC demographically-sampled speech section. Measuring "complexity" in the schoolroom sense of the incidence of subordinate clauses embedded within higher clauses, I found (to my considerable

surprise) that there appears to be a statistically-significant correlation with speakers' age, in the sense that (not just through childhood but on beyond the "critical period" into the thirties, forties, fifties, and sixties) people's speech grows more complex as they get older (Sampson 2001: Chapter 5). If this effect is genuine, it is surely not just fascinating but potentially has implications for social policy and the like.

The proviso "if it is genuine" is important: creating a treebank of casual speech is a time-consuming, expensive business, so the sample available to me was small and the statistical test I applied achieved only a modest level of significance. (Currently I am developing a larger sample, which may in due course establish the finding more robustly – or may show it to have been a meaningless blip.)

If the finding is indeed genuine, because the BNC gives us a snapshot of British speech at one point in history (the early 1990s) it can be interpreted in alternative ways. It might mean that individuals' speech patterns regularly grow syntactically more complex as the individuals age; they always have and they always will. Or it might mean that changes in British society over the twentieth century, perhaps the spread of television and internet use, have led adults born in the 1960s and 1970s to adopt grammatically simpler styles of speech than those which people born in the 1930s adopted at the same age: the younger generation will never come to speak in the way that was natural for their parents.

Syntactic structure is so intimately related to human thought processes that we should surely want to know which of these interpretations is correct. Without corpora, questions like this could never emerge.

I cannot comment to any extent on language teaching, since this is not a topic I know much about. But if it really were the case that the speech of younger Britons is not spontaneously developing the levels of structural complexity found in the speech of previous generations, then one might feel that it should be a priority for primary and secondary education to do what it can to remedy this. Complex speech is not desirable for its own sake; when something can be put simply, that is the best way to put it. But many topics are inherently complicated, and citizens who are capable of engaging with complication in their thinking and speaking will, I would suppose, be better and more fulfilled citizens than those who are forced to oversimplify.

Let me repeat that at present it is far from clear that the correlation of complexity with age is a real phenomenon, let alone which explanation for it is the correct one, if it is real. But a style of linguistics which even potentially leads to consideration of issues like these is surely more worth pursuing than aprioristic theorizing about artificially neat invented examples of language.

References

Chater, N., Reali, F. & Christiansen, M. H. 2009. Restrictions on biological adaptation in language evolution. *Proceedings of the National Academy of Sciences* 106: 1015–1020.

Chomsky, N. 1964. *Current Issues in Linguistic Theory*. The Hague: Mouton.

Chomsky, N. 1980. On cognitive structures and their development: A reply to Piaget. In *Language and Learning*, M. Piattelli-Palmarini (ed.), 35–52. London: Routledge & Kegan Paul.

Edwards, J. 1992. Design principles in the transcription of spoken discourse. In *Directions in Corpus Linguistics*, J. Svartvik (ed.), 129–44. Berlin: Mouton de Gruyter.

Evans, N. & Liberman, S. C. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32: 429–92.

Householder, F. W. 1973. On arguments from asterisks. *Foundations of Language* 10: 365–376.

Kaeding, F. W. 1898. *Häufigkeitswörterbuch der deutschen Sprache*. Steglitz: Privately printed.

Katz, J. J. 1971. *The Underlying Reality of Language and its Philosophical Import*. London: Harper & Row.

Labov, W. 1975. Empirical foundations of linguistic theory. In *The Scope of American Linguistics*, R. Austerlitz (ed.), 77–133. Lisse: Peter de Ridder Press. (Also published separately as *What is a Linguistic Fact?* Lisse: Peter de Ridde Press, 1975).

Pinker, S. 1994. *The Language Instinct*. New York NY: William Morrow.

Pullum, G. K. & Scholz, B. C. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review* 19: 9–50.

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

Sampson, G. R. 1975. Chapter 4 of *The Form of Language*. London: Weidenfeld & Nicolson (Reprinted as Chapter 8 of Sampson, 2001).

Sampson, G. R. 2001. *Empirical Linguistics*. London: Continuum.

Sampson, G. R. 2005. *The 'Language Instinct' Debate,* Rev. edn. London: Continuum.

Sampson, G. R. 2007. Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory* 3: 1–32 & 111–29.

Sampson, G. R. & McCarthy, D. (eds). 2004. *Corpus Linguistics*. London: Continuum.