Language and Linguistics Compass 3 (2009): 1–17, 10.1111/j.1749-818x.2009.00149.x

What is Corpus Linguistics?

Stefan Th. Gries*
University of California

Abstract

Corpus linguistics is one of the fastest-growing methodologies in contemporary linguistics. In a conversational format, this article answers a few questions that corpus linguists regularly face from linguists who have not used corpus-based methods so far. It discusses some of the central assumptions ('formal distributional differences reflect functional differences'), notions (corpora, representativity and balancedness, markup and annotation), and methods of corpus linguistics (frequency lists, concordances, collocations), and discusses a few ways in which the discipline still needs to mature.

At a recent LSA meeting … [with an obvious bow to Frederick Newmeyer]

Question: So, I hear you're a corpus linguist. Interesting, I get to see more and more abstracts and papers and even job ads where experience with corpus-based methods are mentioned, but I actually know only very little about this area. So, what's this all about?

Answer: Yes, it's true, it's really an approach that's gaining more and more prominence in the field. In an editorial of the flagship journal of the discipline, Joseph (2004:382) actually wrote 'we seem to be witnessing as well a shift in the way some linguists find and utilize data – many papers now use corpora as their primary data, and many use internet data'.

Question: My impression exactly. Now, you say 'approach', but that's something I've never really understood. Corpus linguistics – is that a theory or model or a method or what?

Answer: Good question and, as usual, people differ in their opinions. One well-known corpus linguist, for example, considers corpus linguistics – he calls it *computer corpus linguistics* – a 'new philosophical approach […]' Leech (1992:106). Many others, including myself, consider it a method(ology), no more, but also no less (cf. McEnery et al. 2006:7f). However, I don't think this difference would result in many practical differences. Taylor (2008) discusses this issue in more detail, and for an amazingly comprehensive overview of how huge and diverse the field has become, cf. Lüdeling and Kytö (2008, 2009).

Question: Hm … But if you think corpus linguistics is a methodology, …. Well, let me ask you this: usually, linguists try to interpret the data they investigate against the background of some theory. Generative grammarians interpret their acceptability judgments within Government and Binding Theory or the Minimalist Program; some psycholinguists interpret their reaction time data within, for example, a connectionist interactive

activation model – now if corpus linguistics is only a methodology, then what is the theory within which you interpret your findings?

Answer: Again as usual, there's no simple answer to this question; it depends …. There are different perspectives one can take. One is that many corpus linguists would perhaps even say that for them, linguistic theory is not of the same prime importance as it is in, for example, generative approaches. Correspondingly, I think it's fair to say that a large body of corpus-linguistic work has a rather descriptive or applied focus and does actually not involve much linguistic theory.

Another one is that corpus linguistic methods are a method just as acceptability judgments, experimental data, etc. and that linguists of every theoretical persuasion can use corpus data. If a linguist investigates how lexical items become more and more used as grammatical markers in a corpus, then the results are descriptive and/or most likely interpreted within some form of grammaticalization theory. If a linguist studies how German second language learners of English acquire the formation of complex clauses, then he will either just describe what he finds or interpret it within some theory of second language acquisition and so on… .

There's one other, more general way to look at it, though. I can of course not speak for all corpus linguists, but I myself think that a particular kind of linguistic theory is actually particularly compatible with corpus-linguistic methods. These are usage-based cognitive-linguistic theories, and they're compatible with corpus linguistics in several ways. (You'll find some discussion in Schönefeld 1999.) First, the units of language assumed in cognitive linguistics and corpus linguistics are very similar: what is a unit in probably most versions of cognitive linguistics or construction grammar is a symbolic unit or a construction, which is an element that covers morphemes, words, etc. Such symbolic units or constructions are often defined broadly enough to match nearly all of the relevant corpus-linguistic notions (cf. Gries 2008a): collocations, colligations, phraseologisms, …. Lastly, corpus-linguistic analyses are always based on the evaluation of some kind of frequencies, and frequency as well as its supposed mental correlate of cognitive entrenchment is one of several central key explanatory mechanisms within cognitively motivated approaches (cf., e.g. Bybee and Hopper 1997; Barlow and Kemmer 2000; Ellis 2002a,b; Goldberg 2006).

Question: Wait a second – 'corpus-linguistic analyses are *always* based on the evaluation of some kind of frequencies?' What does that mean? I mean, most linguistic research I know is not about frequencies at all – if corpus linguistics is all about frequencies, then what does corpus linguistics have to contribute?

Answer: Well, many corpus linguists would probably not immediately agree to my statement, but I think it's true anyway. There are two things to be clarified here. First, frequency of what? The answer is, there are no meanings, no functions, no concepts in corpora – corpora are (usually text) files and all you can get out of such files is distributional (or quantitative/statistical) information:

- frequencies of occurrence of linguistic elements, i.e. how often morphemes, words, grammatical patterns etc. occur in (parts of) a corpus, etc.; this information is usually represented in so-called *frequency lists*;
- frequencies of co-occurrence of these elements, i.e. how often morphemes occur with particular words, how often particular words occur in a certain grammatical

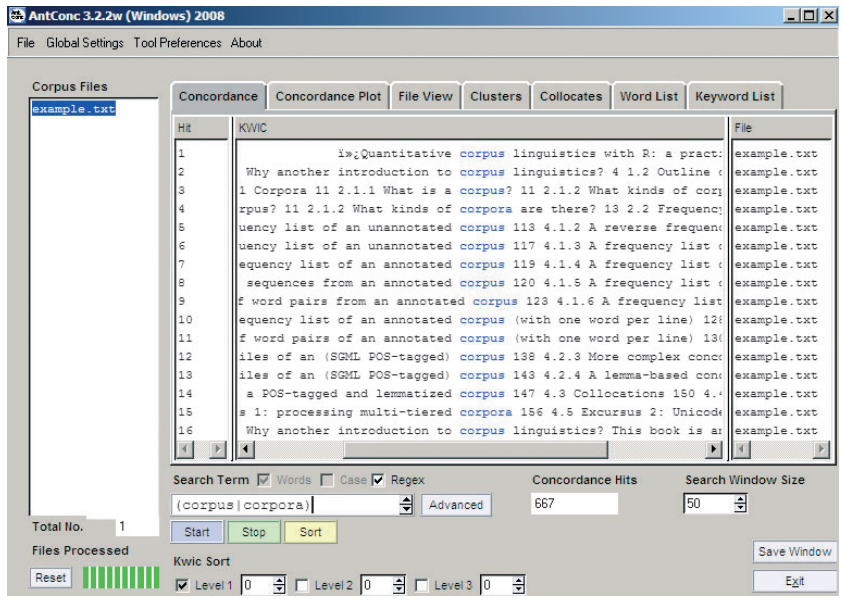

Fig. 1. A concordance output from AntConc 3.2.2w.

construction, etc.; this information is mostly shown in so-called *concordances* in which all occurrences of, say, the word searched for are shown in their respective contexts. Figure 1 is an example.

As a linguist, you don't just want to talk about frequencies or distributional information, which is why corpus linguists must make a particular fundamental assumption or a conceptual leap, from frequencies to the things linguists are interested in, but frequencies is where it all starts. Second, what kind of frequency? The answer is that the notion *frequency* doesn't presuppose that the relevant linguistic phenomenon occurs in a corpus 100 or 1000 times – the notion of *frequency* also includes phenomena that occur only once or not at all. For example, there are statistical methods and models out there that can handle non-occurrence or estimate frequencies of unseen items. Thus, corpus linguistics is concerned with whether

- something (an individual element or the co-occurrence of more than one individual element) is attested in corpora; i.e. whether the observed frequency (of occurrence or co-occurrence) is 0 or larger;
- something is attested in corpora more often than something else; i.e. whether an observed frequency is larger than the observed frequency of something else;
- something is observed more or less often than you would expect by chance [this is a more profound issue than it may seem at first; Stefanowitsch (2006) discusses this in more detail].

This also implies that statistical methods can play a large part in corpus linguistics, but this is one area where I think the discipline must still mature or evolve.

Question: What do you mean?

Answer: Well, this is certainly a matter of debate, but I think that a field that developed in part out of a dissatisfaction concerning methods and data in linguistics ought to be very careful as far as its own methods and data are concerned. It is probably fair to say that many linguists turned to corpus data because they felt there must be more to data collection than researchers intuiting acceptability judgments about what one can say and what one cannot; cf. Labov (1975) and, say, Wasow and Arnold (2005:1485) for discussion and exemplification of the mismatch between the reliability of judgment data by prominent linguists of that time and the importance that was placed on them, as well as McEnery and Wilson (2001: Ch. 1), Sampson (2001: Chs 2, 8, and 10), and the special issue of *Corpus Linguistics and Linguistic Theory* (*CLLT*) 5.1 (2008) on corpus linguistic positions regarding many of Chomsky's claims in general and the method of acceptability judgments in particular. However, since corpus data only provide distributional information in the sense mentioned earlier, this also means that corpus data must be evaluated with tools that have been designed to deal with distributional information and the discipline that provides such tools is statistics. And this is actually completely natural: psychologists and psycholinguists undergo comprehensive training in experimental methods and the statistical tools relevant to these methods so it's only fair that corpus linguists do the same in their domain. After all, it would be kind of a double standard to on the one hand bash many theoretical linguists for their presumably faulty introspective judgment data, but on the other hand then only introspectively eyeball distributions and frequencies. For a while, however, this is exactly what has happened, but the picture is now changing a lot – cf. Mukherjee (2007) for discussion – and more and more corpus linguists use increasingly sophisticated statistical and/or computational methods (cf. papers in Bod et al. 2003). As a result, not only do corpus-linguistic studies become more comprehensive and more precise, the larger degree of quantification also does more justice to language as a multifactorial object of study and makes it easier to relate corpus-based findings to experimental findings. There are now many studies not only in corpus linguistics in particular but also in linguistics in general that combine evidence from these two kinds methods (cf. Kepser and Reis 2005 and *CLLT* 5.1 again).

Question: Ah, ok, well, that makes sense. But even then … one the one hand, it sounds like corpus linguistics should be widely applicable to linguistic problems, but on the other hand, this focus on frequencies and quantification makes it sound as if the applicability of corpus data or the relevance of corpus-linguistic approaches may be severely limited. Or, what is this 'assumption' or 'conceptual leap' that allows you to do more than just distributional number-crunching?

Answer: Oh, on the contrary, corpus data are very widely applicable! Although it's not always openly stated, this assumption underlying most corpus-based analyses is that formal differences reflect, or correspond to, functional differences. Thus, different frequencies of (co-)occurrences of formal elements – again, morphemes, words, syntactic patterns, etc. – are assumed to reflect functional regularities, and 'functional' is understood here in a very broad sense as anything – be it semantic, discourse-pragmatic, … – that is intended to perform a particular communicative function.

Question: I am not sure I get this: how do frequencies reflect functional regularities? Uh …

Answer: Well, one could go back to Bolinger's (1968:127) famous dictum that 'a difference in syntactic form always spells a difference in meaning', the *principle of no synonymy*, Harris's (1970:786) 'difference of meaning correlates with difference of distribution', or Goldberg's (1995:67) more recent version 'if two constructions are syntactically distinct, they must be semantically or pragmatically distinct'. The above assumption is just a little more general than both of these, including any kind of formal distributional difference and any kind of functional aspect.

Consider as an example the case of argument structure, or transitivity alternations such as the 'alternation' between *John sent Mary the book* and *John sent the book to Mary*. Rather than assuming that both syntactic patterns – the ditransitive NP_{AGT} V NP_{REC} NP_{PAT} vs. the prepositional *to*-dative NP_{AGT} V NP_{PAT} $PP_{to\ REC}$ – are functionally identical, another perspective (which is not unique to corpus linguists) might be to assume that, following the principle of no synonymy, the difference in the formal patterns will reflect some functional difference and that corpus-based frequency data can be used to identify that difference. Non-corpus-based studies have suggested, among other things having to do with information structure, that the ditransitive pattern is closely associated with transfer of the patient from the agent to the recipient across a small distance, whereas the prepositional *to*-dative pattern is more associated with transfer over some distance. Now, a corpus-based study by Gries and Stefanowitsch (2004) looked at the statistically preferred verbs in the verb slots of the two patterns and found that the ditransitive's two most strongly preferred verbs are *give* and *tell*, which prototypically involve close proximity of the agent and the recipient, whereas the prepositional dative's two most strongly preferred verbs are *bring* and *play* (as in *he played the ball to him*), which prototypically involve larger distances: if I stand next to you and hand you something, you don't use *bring*, right? These findings are therefore congruent with the proposed semantic difference. Using the same kind of operationalization, i.e. paraphrase of linguistic hypotheses into frequencies, you can investigate an extremely large number of issues. Here's a (necessarily short and biased) list of examples:

- first language acquisition: how often does a child get to hear particular words or patterns in the input and how does that affect the ease/speed with which a child acquires these words and patterns (cf. Goodman et al. 2008 on lexical acquisition, Tomasello 2003 on the acquisition of patterns, and Behrens 2008 for a general recent overview)?
- second/foreign language acquisition: how do we assess second language learners' lexical proficiency (cf. Laufer and Nation 1995; Meara 2005)? how do we determine what linguistic elements to focus on in instruction and how do we use corpora in language teaching (cf. Stevens 1991; Fox 1998; Conrad 2000; Römer 2006)? (For a recent overview of corpus-based methods in SLA, cf. Gries 2008b).
- language and culture: to what degree do frequencies of words reflect differences between cultures (cf. Hofland and Johansson 1982; Leech and Fallon 1992; Oakes and Farrow 2006)?
- historical developments: how can corpora inform language genealogy (cf. Lüdeling 2006)? how do corpus-based methods contribute to research in grammaticalization (cf. Lindquist and Mair 2004 or Hoffmann 2005)?
- phonology: how well can the degree of phonological assimilation or reduction of an expression be predicted on the basis of its components' frequency of co-occurrence (cf. Bybee and Scheibman 1999; Gahl and Garnsey 2004; Ernestus et al. 2006)?
- morphology: what do regular and irregular verb forms reveal about the probabilistic nature of the linguistic system (cf. Baayen and Martín 2005)? how do we assess the productivity of morphological processes (cf. Baayen and Renouf 1996; Plag 1999)?

– syntax: how can we predict which syntactic choices speakers will make (cf. Leech et al. 1994; Wasow 2002; Gries 2003a,b, Bresnan et al. 2007)? what are the overall frequencies of English grammatical structures (cf. Roland et al. 2007)?
– semantics and pragmatics: how do near synonyms differ from each other (cf. Okada 1999; Oh 2000; Gast 2006; Gries and David 2006; Arppe and Järvikivi 2007; Divjak forthcoming)? how are antonyms acquired (cf. Jones and Murphy 2005)? how can we approach complex multifactorial notions such as idiomaticity and compositionality (cf. Barkema 1993; Langlotz 2006; Wulff 2008)? how come that some words such as *happen* and *set in* have a negative twang to them (cf. Whitsitt 2005; Dilts and Newman 2006; Bednarek 2008)?
– plus applications in psycholinguistics (e.g. syntactic priming/persistence), stylistics, sociolinguistics, forensic linguistics (e.g. authorship attribution), etc.

Note also that, frequency information *per se* is often a crucial factor to control in psycholinguistic experiments as frequencies of occurrence are correlated with, among other things, reaction times.

Question: Wow, ok, that's certainly more diverse than I would've thought. From what I had seen, I thought most corpus-based work is purely descriptive and maybe lexicographic or applied in nature. Also, I thought that many people would now use the World Wide Web as a corpus – I now often read something like 'a web search revealed that X is more frequent than Y …', and Joseph's comment you mentioned earlier suggests the same – but then many of the applications you mention don't sound like as if that can be true, or can it?

Answer: Sigh … well, yes and no. Yes, it's true that a growing number of linguists now often query a search engine to, say, determine frequencies of words or patterns. Unfortunately, this practice can result in quite some problems, some of which are technical in nature while others are more theoretical. As for the technical problems, it's well-known by now that the frequencies returned by Google, Yahoo, and other search engines are very variable and may, thus, be unreliable, and web data come with a variety of other problems, too; the special issue 29.3 (2003) of *Computational Linguistics*, Hundt et al. (2007), and Eu (2008) would be good places to read on this.

Question: Yes, I actually heard about the technical problems, but what are the theoretical problems? And then I still want to know what exactly a corpus is! Is that just any collection of '(text) files'? And how do you access it? I guess nowadays this is all done computationally?

Answer: Feel free to go to London and manually browse the index cards of the Survey of English Usage (http://www.ucl.ac.uk/english-usage/), one of the earliest corpora. Joking aside, yes, nowadays, corpus-linguistic studies are nearly always done computationally as virtually all corpora are text collections stored in the form of plain ASCII or, increasingly commonly, Unicode text files that can be loaded, manipulated, and processed platform-independently. This doesn't mean, however, that corpus linguists only deal with raw text files; on the contrary, some corpora come with linked audio files or are shipped with sophisticated retrieval software that makes it possible to retrieve immediately the position of an utterance in an audio file or look for precisely defined syntactic and/or lexical patterns etc. It does mean, however, that you would have a hard time finding corpora on

paper, in the form of punch cards or digitally in proprietary binary formats (such as Microsoft Word's DOC format); the current standard is probably text files with XML annotation (cf. McEnery et al. 2006:28, 35, and passim as well as the website of the Text Encoding Initiative at http://www.tei-c.org/Guidelines/P4/html/SG.html).

But, let me come back to the question about the theoretical problems, which will actually also lead to what at least *I* think a corpus is. These problems are concerned with a general problem of scientific inquiry. When you study any phenomenon, let's assume a linguistic phenomenon such as the frequencies of words or syntactic constructions, or the productivity of a particular word-formation process or whatever, you usually want to be able to generalize from the data you investigate to, in our cases, the language as a whole or a particular register or genre. Now the issue is, can one generalize from the web or what can one generalize to?

Question: Sure, the web contains *all sorts of* information… .

Answer: … actually, it's not quite that certain. Yes, the web contains all sorts of information, especially now in the web 2.0 era with community forums, personal webspaces, blogs, etc. But think about it: the question is, what does the web not contain, what does it contain, and how many of the different things there are does it contain? For example, what are contents that you won't find on the web? Yes, people post very many private things on their sites, but, for instance, personal diaries, intimate conversations, love letters, or other private letters, confessions by criminals, jury transcripts, the conversation between a driving instructor and his student, the language exchanged during a gang brawl may be important for some linguistic studies, but not found on the web (at least not often). True, many of these you will also not find in corpora, but this just goes to show that the web does not contain everything. Also, think about contents you *do* find on the web a lot: marketing and advertising, computer and tech language – just google *Java* or *Ruby* and note how often the island or the gem show up in the first 50 hits – journalese, pornography, scientific texts etc. It's not at all obvious how well you can generalize from the web.

Question: Ok, ok, I get it. But then how are corpora that do not only contain texts from the web better?

Answer: Now, many corpora have usually been put together with an eye to taking care of these issues, to make it possible to generalize from a corpus to a language as a whole or at least to a particular variety, register etc. Thus, corpus compilers usually try to make their corpora representative and balanced. By *representative*, I mean that the different parts of the linguistic variety I'm interested in are all manifested in the corpus. For example, if I was interested in phonological reduction patterns of speech of Californian adolescents and recorded only parts of their conversations with several people from their peer group, my corpus would not be representative in the above sense because it would not reflect the fact that some sizable proportion of the speech of Californian adolescents may also consist of dialogs with a parent, a teacher, etc., which would optimally also have to be included. By *balanced*, I mean that ideally not only should all parts of which a variety consists be sampled into the corpus but also that the proportion with which a particular part is represented in a corpus should reflect the proportion the part makes up in this variety and/or the importance of the part in this variety. For example, if I know that dialogs make up 65% of the speech of Californian adolescents, approx. 65% of my corpus should consist of dialog recordings.

Question: But that's ridiculous: we don't know these percentages!

Answer: Sadly enough, that's true – we don't. These criteria are admittedly more of a theoretical ideal. Even if resources were no problem, how would we measure the proportion that dialogs make up of the speech of Californian adolescents? We can only record a tiny sample of all Californian adolescents, and how would we measure the proportion of dialogs – in terms of time? in terms of sentences? in terms of words? And if we tried to compile a corpus representative of a language as a whole, then how would we measure the importance of a particular linguistic variety?

Question: Well, conversational speech is primary, but I wouldn't know what the next most important variety is.

Answer: I suppose that's the most widely held belief, and corpus compilers often aim at including as much spoken language as possible because, as you say yourself, spoken language, especially spoken conversation, is typically considered the most basic form of language use. On the other hand, a single catchy and thus salient newspaper headline read by millions of people may have a much larger influence on every reader's linguistic system and on the language 'as a whole' than twenty hours of dialog as usual. Anyway, representative and balanced corpora are a theoretical ideal corpus compilers constantly bear in mind, but the ultimate and exact way of compiling a truly representative and balanced corpus has eluded us so far. If you want to read on this, Biber's (1990, 1993) work is most instructive, and McEnery et al. (2006: Sections A2, A8, and B1) provide good summary sections.

Question: I see.

Answer: There's one final criterion for something to be a corpus. The texts that make up the corpus must have been produced in a natural communicative setting. That means that the texts were spoken or written for some authentic communicative purpose, but not for the purpose of putting them into a corpus. For example, many corpora consist to a large degree of newspaper articles. These are of course often included for convenience's sake, but they also meet the criterion of having been produced in a natural setting because journalists write the article to be published in newspapers and magazines and to communicate something to their readers, but not because they want to fill a linguist's corpus. (On the other hand, journalese is often heavily edited.) Similarly, if I obtained permission to record all of a particular person's conversations in one week, then hopefully, while the person and his interlocutors usually are aware of their conversation being recorded, I will obtain authentic conversations rather than conversations produced only for the sake of my corpus.

Question: Sounds like there should be a huge variety of different corpora… .

Answer: Absolutely. In fact, you should know that corpora differ in a variety of ways. There are a few distinctions you should be familiar with if only to be able to find the right corpus for what you want to investigate. The most basic distinction is that between *general corpora* and *specific corpora*. The former intend to be representative and balanced for a language as a whole – within the above-mentioned limits, that is – while the latter are by design restricted to a particular variety, register, genre, …. Then, there's a difference between *diachronic corpora* and *synchronic corpora*. The former aim at representing how a

language/variety changes over time while the latter provide, so to speak, a snapshot of a language/variety at one particular point of time (which may well be a decade). Yet another distinction is that between *monolingual corpora* and *parallel corpora*. As you may already guess from the names, the former have been compiled to provide information about one particular language/variety etc., whereas the latter ideally provide the same text in several different languages. Examples include translations from EU Parliament debates into the 23 languages of the European Union or the Canadian Hansard corpus (http://www.isi.edu/natural-language/download/hansard/), containing Canadian Parliament debates in English and French. Again ideally, a parallel corpus doesn't just have the translations in the *n* different languages, but the translations are also sentence-aligned, such that for every sentence in language L_1, you can automatically retrieve its translation in the languages L_2 to L_n. The final distinction to be mentioned here is that of *static corpora* vs. *dynamic/monitor corpora*. Static corpora have a fixed size, whereas dynamic corpora do not as they may be constantly extended with new material. There really are a lot of interesting issues involved here. If you ever think of compiling your own little corpus, the papers in Wynne (2005) and Beal et al. (2007a,b) are great references in addition to the ones I already mentioned.

Question: Ok, now I at least understand *what* it is you can find in corpora. So, but *how* do you then get the words out of the corpus – with a text editor? Well, probably not a text editor because you said something about not using Word… .

Answer: Well, in a way your question is a bit too narrow because there is more to get out of corpora than just words. This is true in two ways: First, it's true in the trivial sense that corpora don't just contain words, but of course words in syntactic patterns and, in corpora with audio files, words in syntactic patterns in turns with particular intonation curves, etc. – after all, we said that a corpus contains language as used in authentic communicative situations, and in such situations we don't just string words together randomly. Second, it's true in the much less trivial sense that many corpora contain more than just words in their syntactic constructions: they also contain additional information. First, corpus files often contain so-called *metadata*, i.e. information that identifies the source(s) of the data in the corpus file, for example, in which newspaper and when the article in the corpus file was published or the number of interlocutors whose recorded conversation the file contains as well as sociological or sociolinguistic information about them (e.g. their age, profession, and highest degree). Also, markup may contain copyright information, which may be relevant because not all corpora may be universally available. For example, for copyright reasons, the first version of the British National Corpus wasn't available in the USA – only the later British National Corpus World edition, which didn't contain some of the files of the first version, became available in the USA. Such information is often stored at the beginning of the corpus files in a so-called header, a part that precedes the actual text, the actual recording, etc.

Second, corpus files often contain what is sometimes referred to as *markup*, information about what the physical appearance of the original documents was like: the document structure, headings, paragraph breaks, etc.

Question: That makes sense. But I thought this is what you meant by *annotation* earlier… .

Answer: Well, unfortunately, the terms *markup* and *annotation* are not used the same way by everybody; cf. Garside et al. (1997), Bird and Liberman (2001), Bowker and Pearson (2002: Ch. 5), and Leech (2005) for much input. What is often meant by *annotation*,

however, is interpretative linguistic information. That is, *raw corpora* consist of files only containing the corpus material [cf. (1)] (and maybe metadata and markup), i.e. largely objective information. *Annotated corpora*, on the other hand, also contain information about the language data in the corpus part, information that represents a particular linguistic analysis. Thus, annotation is often debatable, because not everyone may subscribe to the theory or model on which the annotation is based.

Question: 'Represents a particular kind of analysis' – what does that mean?

Answer: One relatively uncontroversial example of annotation is *lemmatization*. A corpus is lemmatized when, for example, each word in the corpus is followed by its lemma, i.e. the form under which you would look it up in a dictionary [cf. (2)]. A much less uncontroversial form of annotation involves so-called *part-of speech (POS) tags*, where each word in the corpus is preceded or followed by an abbreviation giving the word's part of speech and sometimes also some morphological information [cf. (3)]. Here, different human analysts (and computer algorithms) may well arrive at conflicting and/or incorrect results. It gets even more difficult (in the sense of arriving at accurate and unanimous annotation) once we turn to syntactic annotation: a corpus may be syntactically parsed, i.e. each word comes with information about where in the sentence's/utterance's syntactic tree it is located [cf. (4)]; such corpora are often referred to as treebanks; the Penn Treebank project (cf. http://www.cis.upenn.edu/~treebank/), the British Component of the International Corpus of English (ICE; cf. http://www.ucl.ac.uk/english-usage/ice/index.htm), and the German Negra and TIGER corpora (cf. http://www.coli.uni-sb.de/sfb378/negra-corpus/ and http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGER-Corpus/) are well-known examples, and Abeillé (2003) a well-known reference work. A corpus may also be phonologically annotated [cf. (5)]. Finally, a corpus may contain several different annotations on different lines (or tiers) at the same time, a format known as CHAT format, which is especially common in language acquisition corpora [cf. (6)].

(1) I did get a postcard from him.

(2) I<I> did<do> get<get> a<a> postcard<postcard> from<from> him<he>.<punct>

(3) I<PersPron> did<VerbPast> get<VerbInf> a<Det> postcard<NounSing>
from<Prep> him<PersPron>. <punct>

(4)
```
<Subject, NP>
        I<PersPron>
<Predicate, VP>
        did<Verb>
        get<Verb>
        <DirObject, NP>
            a<Det>
            postcard<NounSing>
        <Adverbial, PP>
            from<Prep>
            him<PersPron>.
```

(5) [@:] I ^did get a !p\ostcard fr/om him# - -

(6) *CHI: I did get a postcard from him
%mor: pro|I v|do&PAST v|get det|a n|postcard prep|from pro|him.
%lex: get
%syn: trans

There are also corpora – most notably perhaps the American National Corpus (http://www.anc.org/) – with standoff annotation, that is, a format where the data and the markup and annotation are in separate but cross-referenced files, but this is more difficult to exemplify here (cf. McEnery et al. 2006: Sections A3 and A4). Thus, you can often get much more out of a corpus than just words.

Now that that's clarified, back to your question of how we get information out of corpora. Basically, there are several possibilities. First, there are many free corpus programs out there which come with relatively easy to use point-and-click graphical user interfaces and provide, for example, the kind of output shown in Figure 1; freely available programs include AConcorde, AntConc, ConcApp, Concorder, Corpus Wizard, Simple Concordance Program, TextStat, Xaira, Multilingual Concordancer, Corpus Search 2, etc. All of these programs can be applied to standard text formats, where by 'standard text format' I mean files which look like (1) to (3) above. Second, as you have seen, some corpora come in formats which pretty much require the retrieval software they come with. For example, the ICE-GB corpus or the Penn Treebank are fully tagged and parsed and the resulting organization of the text files is similar to, but much more complex than, (4) above and cannot be handled well with the above popular programs, but require specialized software. Third, you can use search (and replace) software (cf. http://www.thefreecountry.com/programming/searchandreplace.shtml for examples) on most of the standard text formats. Fourth, there's a growing number of corpus linguists who use programming languages to do their data processing. The languages used most are probably Perl (http://www.perl.org/), Python (http://www.python.org/), and then maybe R, the open source programming language derived from S-Plus (http://www.r-project.org/). All three are freely available programming exhibiting object-oriented, imperative, and functional characteristics (cf. http://en.wikipedia.org/wiki/Comparison_of_programming_languages for a comparison); Perl, Python, and R are probably best-known for text-processing capabilities, scripting for web applications, and statistics / statistical graphics respectively, but all three are in fact multi-purpose languages.

Question: Well, why would anyone *not* want to use a program with a nice point-and-click GUI, but a programming language?

Answer: Well, again, there are several reasons. As I mentioned, not all corpus programs can handle all sorts of files so sometimes you need a more flexible tool. Then, some corpus programs – none of the ones I mentioned, though – are commercial programs while the programming languages are freely available (and provide the added bonus that you get to learn how to program). Then, when you use pre-configured corpus programs, you're a little bit at the mercy of the company or individual selling them: they may discontinue development or sales, they may change crucial parameters of how the programs work, they …

Question: Wait, wait, what do you mean, 'crucial parameters of how they work'?

Answer: Well, for instance, consider frequency lists, i.e. lists that tell you which words occur in a corpus or a part of a corpus and how often each occurs. Now such lists depend very much on what you think a word is and remember we're now talking about a computer knowing what a word is. The computer only processes strings of characters and must therefore identify all word tokens in the file(s), a process called *tokenization*. So, you might want to define to a computer what a word as 'a string of letters surrounded by spaces'. Well, that does the trick most of the time, but sometimes a word is not followed by a space, but by a punctuation mark. Ok, so you might say a word is 'a string of letters surrounded by spaces or punctuation marks'. Then what about *Ph.D.* or just *Dr.*? Is a bracket a punctuation mark? What about hyphens? Do we want to say *ill-defined* is two words? How do we handle spelling variation? Do we want to be able to say that *armchair-linguist* and *armchair linguist* are the same words? And *favour* and *favor*? Is *John's book* two or three words? If you think *John's book* is two words, you might then be tempted to say a word is 'a string of letters and/or hyphens and/or apostrophes surrounded by spaces or punctuation marks.' But then what about *he's going* or *isn't it*? Aren't these three words each? And if you say 'yes', then note that the apostrophe in *isn't it* is not even where we would split *isn't it* up into words: we would say the 'words' are 'is', 'n't', and 'it', not 'isn', 't', and 'it'. Let me make it worse for you: isn't *1960* a word, and how many words does *a 25-year-old man* contain, or a link such as *http://www.linguistics.ucsb.edu*? I guess it's obvious by now: it's far from clear how to define to a computer what a word is. And I haven't even mentioned the issues that arise when you look at languages other than English – how do you treat a Russian expression in Cyrillic fonts such as 'Чайковский' in an American newspaper text? Programmers make different distinctions and as a result different corpus software will output different frequency lists for the same corpus, and a corpus program whose definition was changed will output a frequency list for a corpus that is different from a frequency list for the same corpus made with an earlier version of the same program.

Question: I see …

Answers: There are other things, too. Recall that I mentioned that corpus files often have a header that contains information about the file, right? Now, if you want to create a frequency list of a corpus file, then you of course do *not* want to include the header in that count, but many ready-made programs, even commercial ones, can't distinguish the header from the rest of the corpus and will simply count all words in the header, too. And imagine a corpus with part-of-speech tags: can the program handle those properly?

Question: Hm …

Answer: One final big advantage of programming languages, therefore, is that you are in the driver's seat. *You* define what a word is and can make sure that your data do not depend on how other people's definitions change. Second, often your specific application requires functionality that a ready-made corpus program does not offer. For example, you may want to retrieve all verb forms (e.g. *go*, *goes*, *going*, *gonna* – I'm sure you thought of that one – *went*, and *gone*) of 120 verb lemmas (e.g. the lemma *to go*), and store each occurrence of any of these six forms with the preceding and following ten words in a particular sorting order into a separate file for each lemma (so that you can look at each lemma's file separately later). I know of no concordance program that can do this. With a programming language, this can be relatively easy, and there are many similar and many

much more complex scenarios where the same is true. I for myself used Perl, but have now begun to rely more on the programming language R, which is not as fast as Perl, but just as powerful in terms of search facilities, it's much simpler to program, unlike many commercial programs it can handle Unicode (http://unicode.org/) and XML (http://www.xml.com/), allows you to interface with SQL databases (cf. Davies 2005 for recent exciting advances to make corpora available as databases) and even allows you to do statistical and graphical exploration and evaluation within one environment (cf. Gries 2009). I mentioned earlier the need for corpus linguistics to take statistical methods more seriously – well, to my mind at least, R is a tool that integrates nearly everything most corpus linguists need to do.

Question: That sure sounds great. On the other hand, it makes me wonder …. Many of the things you mentioned sound like you need huge corpora before you can do anything let alone statistics. I would think that there are no large corpora for 99% of the world's languages so where does that leave corpus linguistics?

Answer: You are absolutely right: this is a big issue and another area in which corpus linguistics still has a lot to do. For English, esp. British and American English, many corpora are available: among the English-language corpora that include spoken and written language, the British National Corpus (cf. http://www.natcorp.ox.ac.uk/), the Bank of English (http://www.collins.co.uk/books.aspx?group=153), and the International Corpus of English (cf. above) with several varieties of English are probably among the currently most widely used corpora, and we're eagerly awaiting the release of the complete American National Corpus, using the BYU Corpus of American English (http://www.americancorpus.org/) in the meantime. Among English-language corpora that contain only written language, the Brown corpus (written American English of the 1960s) and its counterpart, the LOB corpus (written British English of the 1960s) as well as the more contemporary counterparts Frown and Flob are very widely used (for these four, cf. http://icame.uib.no/). But as I said, for English, there are very many corpora out there; for example, English is the language with, as far as I can see, the largest amount of corpus data on diachronic development (e.g. the Helsinki corpus; cf. http://icame.uib.no/), first language acquisition (e.g. the CHILDES archive), second/foreign language learning (e.g. the International Corpus of Learner English; cf. http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm), etc. No other language is that well documented… .

Then, for many of the usual suspects, i.e. many of the typical Indo-European languages, we also have some corpora available: following the British National Corpus, many languages now have their own national corpus. Such national corpora are now available for, say, Croatian, Czech, Greek, Hungarian, and Russian, but other corpora such as the Corpus del Espanol (for Spanish), the La Repubblica Corpus (for Italian), the IPI PAN corpus of Polish, the Cosmas II corpus collection for German etc. are useful resources. Unfortunately, some of these corpora contain little or no spoken language, are only available via the WWW or with specialized software running only on one operation system, which sometimes limits their usability considerably.

Now, for most other languages, the situation is much more gloomy or even dramatic because there are no corpora at all, only small corpora, a few texts, a few recorded conversations or collections of narratives in the possession of individual researchers, departments, or indigenous communities. It goes without saying that this doesn't make it any easier to conduct corpus-based research on these languages and that there's a great need for compiling at least small text databases for highly endangered languages. Given the

kinds and small amounts of material that are available for such languages, issues of balancedness and representativity are not even among the most problematic questions, but issues of how to annotate the data, standardize markup and annotation, make resources compatible and interoperable with each other etc. are pressing concerns to which corpus linguists, field linguists, and language documentation experts must now work on together; in an ideal world resources and technologies from different projects such as, say, the World Atlas of Linguistic Structures (Haspelmath et al. 2005), the Electronic Metastructure for Endangered Languages Data project, and the Open Language Archives Community would be combined. You can find a very comprehensive overview of existing corpora and software at David Lee's excellent site at http://devoted.to/corpora.

However, the good news is, first, that many linguists are of course aware of this and a great deal of work is being carried out to make data on many different underdocumented languages available. Second, depending on what one wants to study, sometimes even small corpora (of whatever language) can already go a long way. For example, Hollmann and Siewierska (2006) study ditransitive constructions in comparatively small corpora of Lancashire dialects, Berez and Gries (forthcoming) is a multifactorial study of middle voice marking in 'only' 2800 lines of Dena'ina (Athabaskan, Alaska), and Ghadessy et al. (2001) is a whole volume on how small corpora can be used in the context of English language teaching. So, while larger corpora are often better and certainly desirable, even with corpora whose sizes are limited to what is often the maximum size for underdocumented languages, interesting studies can be undertaken.

Question: Ok, good to know. Well, thanks a lot for all the info, I'll certainly look out for more corpus-based work. Nice talking to you.

Answer: Nice talking to you, too, see you around.

Short Biography

Stefan Th. Gries is a quantitative corpus linguist at the intersection of corpus linguistics and computational linguistics, who uses a variety of methods to investigate linguistic topics such as morpho-phonology (the formation of morphological blends), syntax (syntactic alternations), the syntax-lexis interface (collostructional analysis), and semantics (polysemy, antonymy, and near synonymy in English and Russian) and corpus-linguistic methodology (corpus homogeneity and comparisons, dispersion measures, and other quantitative methods). He is a cognitively oriented linguist in the wider sense of seeking linguistic explanations in terms of cognitive processes. He has published three books – one research monograph, an introduction to statistics with R for linguists, and a book on corpus linguistics with R – and articles in *Cognitive Linguistics*, *International Journal of Corpus Linguistics*, *Linguistics*, *Journal of Psycholinguistic Research*, *Corpora*, *Annual Review of Cognitive Linguistics*, *The Mental Lexicon*, *Literary and Linguistic Computing*, *Journal of Quantitative Linguistics*, *Journal of Child Language*, and others). He is founding Editor-in-Chief of the international peer-reviewed journal *Corpus Linguistics and Linguistic Theory* and performs editorial functions for *Cognitive Linguistics*, *Constructions and Frames*, *Language and Cognition*, and *CogniTextes*.

Note

* Correspondence address: Stefan Th. Gries; Department of Linguistics; University of California, Santa Barbara; Santa Barbara, CA 93106-3100, USA. E-mail: stgries@gmail.com.

Works Cited

Abeillé, Anne (ed.) 2003. Treebanks: building and using parsed corpora. Dordrecht: Kluwer Academic Publishers.

Arppe, Antti, and Juhani Järvikivi. 2007. Every method counts: combining corpus-based and experimental evidence in the study of synonymy. Corpus Linguistics and Linguistic Theory 3.131–59.

Baayen, R. Harald, and Fermín Moscoso del Prado Martín. 2005. Semantic density and past-tense formation in three Germanic languages. Language 81.666–98.

——, and Antoinette Renouf. 1996. Chronicling The Times: productive lexical innovations in an English newspaper. Language 72.69–96.

Barkema, Henk. 1993. Idiomaticity in English NPs. English language corpora: design, analysis, and exploitation, ed. by Jan Aarts, Pieter de Haan and Nelleke Oostdijk, 257–78. Amsterdam: Rodopi.

Barlow, Michael, and Suzanne Kemmer (eds) 2000. Usage-based models of language. Stanford, CA: CSLI Publications.

Beal, Joan C., Karen P. Corrigan, and Hermann L. Moisl (eds) 2007a. Creating and digitizing language corpora: synchronic databases, Vol. 1. Basingstoke: Palgrave MacMillan.

——. (eds) 2007b. Creating and digitizing language corpora: diachronic databases, Vol. 2. Basingstoke: Palgrave MacMillan.

Bednarek, Monika. 2008. Semantic preference and semantic prosody re-examined. Corpus Linguistics and Linguistic Theory 4(2).119–40.

Behrens, Heike (ed.) 2008. Corpora in language acquisition research: history, methods, perspectives. Amsterdam, Philadelphia: John Benjamins.

Berez, Andrea L., and Stefan Th. Gries. forthcoming. Correlates to middle marking in Dena'ina iterative verbs. International Journal of American Linguistics.

Biber, Douglas. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. Literary and Linguistic Computing 5.257–69.

Biber, Douglas. 1993. Representativeness in corpus design. Literary and Linguistic Computing 8.243–57.

Bird, Steven, and Mark Liberman. 2001. A formal framework for linguistic annotation. Speech Communication 33.23–60.

Bod, Rens, Jennifer Hay, and Stephanie Jannedy (eds) 2003. Probabilistic linguistics. Cambridge, MA: The MIT Press.

Bolinger, Dwight L. 1968. Entailment and the meaning of structures. Glossa 2.119–27.

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

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. Cognitive foundations of interpretation, ed. by G. Boume, I. Kraemer and J. Zwarts, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bybee, Joan L., and Paul J. Hopper (eds) 1997. Frequency and the emergence of linguistic structure. Amsterdam, Philadelphia: John Benjamins.

Bybee, Joan, and Joanne Scheibman. 1999. The effect of usage on degrees of constituency: the reduction of *don't* in English. Linguistics 37.575–96.

Conrad, Susan. 2000. Will corpus linguistics revolutionize grammar teaching in the 21st century? TESOL Quarterly 34.548–60.

Davies, Mark. 2005. The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation. International Journal of Corpus Linguistics 10.301–28.

Dilts, Philip, and John Newman. 2006. A note on quantifying 'good' and 'bad' prosodies. Corpus Linguistics and Linguistic Theory 2.233–42.

Divjak, Dagmar S. forthcoming. Structuring the lexicon: a clustered model for near-synonymy. Berlin, New York: Mouton de Gruyter.

Ellis, Nick C. 2002a. Frequency effects in language processing and acquisition. Studies in Second Language Acquisition 24.143–88.

——. 2002b. Reflections on frequency effects in language processing. Studies in Second Language Acquisition 24.297–339.

Ernestus, Mirjam, Mybeth Lahey, and R. Harald Baayen. 2006. Lexical frequency and voice assimilation. Journal of the Acoustical Society of America 120.1040–51.

Eu, Jinseung. 2008. Testing search engine frequencies: patterns of inconsistency. Corpus Linguistics and Linguistic Theory 4.177–208.

Fox, Gwyneth. 1998. Using corpus data in the classroom. Materials development in language teaching, ed. by Brian Tomlinson, 25–43. Cambridge: Cambridge University Press.

Gahl, Susanne, and Susan Marie Garnsey. 2004. Knowledge of grammar, knowledge of usage: syntactic probabilities affect pronunciation variation. Language 80.748–75.

Garside, Roger, Geoffrey Leech, and Anthony McEnery (eds) 1997. Corpus annotation: linguistic information from computer text corpora. London, New York: Longman.

Gast, Volker. 2006. The distribution of *also* and *too*: a preliminary corpus study. Zeitschrift für Anglistik und Amerikanistik 54.163–76.
Ghadessy, Mohsen, Alex Henry, and Robert L. Roseberry (eds) 2001. Small corpus studies and ELT. Amsterdam, Philadelphia: John Benjamins.
Goldberg, Adele E. 1995. Constructions: a construction grammar approach to argument structure. Chicago, IL: The University of Chicago Press.
Goldberg, Adele E. 2006. Constructions at work. Oxford: Oxford University Press.
Goodman, Judith C., Philip S. Dale, and Ping Li. 2008. Does frequency count? Parental input and the acquisition of vocabulary. Journal of Child Language 35.515–31.
Gries, Stefan Th. 2003a. Multifactorial analysis in corpus linguistics: a study of particle placement. London, New York: Continuum Press.
——. 2003b. Towards a corpus-based identification of prototypical instances of constructions. Annual Review of Cognitive Linguistics 1.1–27.
——. 2008a. Phraseology and linguistic theory: a brief survey. Phraseology: an interdisciplinary perspective, ed. by Sylviane Granger and Fanny Meunier, 3–25. Amsterdam, Philadelphia: John Benjamins.
——. 2008b. Corpus-based methods in analyses of SLA data. Handbook of cognitive linguistics and second language acquisition, ed. by. Peter Robinson and Nick C. Ellis, 406–31. New York: Routledge, Taylor and Francis Group.
——. 2009. Quantitative corpus linguistics with R: a practical introduction. New York: Routledge.
Gries, Stefan Th., and Caroline V. David. 2006. This is kind of/sort of interesting: variation in hedging in English. Towards multimedia in corpus linguistics. Studies in variation, contacts and change in English 2, ed. by. Pahta Päivi, Irma Taavitsainen, Terttu Nevalainen and Jukka Tyrkkoö. Helsinki: University of Helsinki.
——, and Anatol Stefanowitsch. 2004. Extending collostructional analysis: a corpus-based perspective on 'alternations'. International Journal of Corpus Linguistics 9.97–129.
Harris, Zellig S. 1970. Papers in structural and transformational linguistics. Dordrecht: Reidel.
Haspelmath, Martin, Matthew Dryer, David Gil, and Bernard Comrie (eds) 2005. The world atlas of language structures. Oxford: Oxford University Press.
Hoffmann, Sebastian. 2005. Grammaticalization and English complex prepositions. London, New York: Routledge.
Hofland, Knut, and Stig Johansson. 1982. Word frequencies in British and American English. London: Longman.
Hollmann, Willem, and Anna Siewierska. 2006. Corpora and (the need for) other methods in a study of Lancashire dialect. Zeitschrift für Anglistik und Amerikanistik 54.203–16.
Hundt, Marianne, Nadja Nesselhauf, and Carolin Biewer (eds) 2007. Corpus linguistics and the web. Amsterdam: Rodopi.
Jones, Steven, and M. Lynne Murphy. 2005. Using corpora to investigate antonym acquisition. International Journal of Corpus Linguistics 10.401–22.
Joseph, Brian. 2004. On change in *Language* and change in language. Language 80.381–3.
Kepser, Stephan, and Marga Reis (eds) 2005. Linguistic evidence: empirical, theoretical and computational perspectives. Berlin, New York: Mouton de Gruyter.
Labov, William. 1975. Empirical foundations of linguistic theory. The scope of American linguistics, ed. by Robert Austerlitz, 77–133. Lisse: The Peter de Ridder Press.
Langlotz, Andreas. 2006. Idiomatic creativity: a cognitive-linguistic model of idiom-representation and idiom-variation in English. Amsterdam, Philadelphia: John Benjamins.
Laufer, Batia, and Paul Nation. 1995. Vocabulary size and use: lexical richness in L2 written production. Applied Linguistics 16.307–22.
Leech, Geoffrey N. 1992. Corpora and theories of linguistic performance. Directions in corpus linguistics. Proceedings of Nobel Symposium 82, ed. by Jan Svartvik, 105–22. Berlin, New York: Mouton de Gruyter.
——. 2005. Adding linguistic annotation. Developing linguistic corpora: a guide to good practice, ed. by Martin Wynne, 17–29. Oxford: Oxbow Books <http://ahds.ac.uk/linguistic-corpora/>.
Leech, Geoffrey N., and Roger Fallon. 1992. Computer corpora: what do they tell us about culture? ICAME Journal 16.29–50.
——, Brian Francis, and Xunfeng Xu. 1994. The use of computer corpora in the textual demonstrability of gradience in linguistic categories. Continuity in linguistic semantics, ed. by Catherine Fuchs and Bernard Victorri, 57–76. Amsterdam, Philadelphia: John Benjamins.
Lindquist, Hans, and Christian Mair (eds) 2004. Corpus approaches to grammaticalization in English. Amsterdam, Philadelphia: John Benjamins.
Lüdeling, Anke. 2006. Using corpora in the calculation of language relationships. Zeitschrift für Anglistik und Amerikanistik 54.217–27.
——, and Merja Kytö. 2008. Corpus linguistics: an international handbook, Vol. 1. Berlin, New York: Mouton de Gruyter.
——. 2009. Corpus linguistics: an international handbook, Vol. 2. Berlin, New York: Mouton de Gruyter.
McEnery, Tony, and Andrew Wilson. 2001. Corpus linguistics, 2nd edn. Edinburgh: Edinburgh University Press.

McEnery, Anthony, Richard Xiao, and Yukio Tono. 2006. Corpus-based language studies: an advanced resource book. London, New York: Routledge.
Meara, Paul. 2005. Lexical frequency profiles: a Monte Carlo analysis. Applied Linguistics 26.32–47.
Mukherjee, Joybrato. 2007. Corpus linguistics and linguistic theory: general nouns and general issues. International Journal of Corpus Linguistics 12.131–47.
Oakes, Michael P., and Malcolm Farrow. 2006. Use of the chi-squared test to examine vocabulary differences in English language corpora representing seven different countries. Literary and Linguistic Computing 22.85–99.
Oh, Sun-Young. 2000. *Actually* and in *fact* in American English: a data-based analysis. English Language and Linguistics 4.243–68.
Okada, Sadayuki. 1999. On the function and distribution of the modifiers *respective* and *respectively*. Linguistics 37.871–903.
Plag, Ingo. 1999. Morphological productivity. Berlin, New York: Mouton de Gruyter.
Roland, Douglas, Frederic Dick, and Jeffrey L. Elman. 2007. Frequency of basic English grammatical structures: a corpus analysis. Journal of Memory and Language 57.348–79.
Römer, Ute. 2006. Pedagogical applications of corpora: some reflections on the current scope and a wish list for future developments. Zeitschrift für Anglistik und Amerikanistik 54.121–34.
Sampson, Geoffrey. 2001. Empirical linguistics. London, New York: Continuum Press.
Schönefeld, Doris. 1999. Corpus linguistics and cognitivism. International Journal of Corpus Linguistics 4.137–71.
Stefanowitsch, Anatol. 2006. New York, Dayton (Ohio), and the raw frequency fallacy. Corpus Linguistics and Linguistic Theory 1.295–301.
Stevens, Vance. 1991. Classroom concordancing: vocabulary materials derived from relevant, authentic text. English for Specific Purposes 10.35–46.
Taylor, Charlotte. 2008. What is corpus linguistics? What the data says. ICAME Journal 32.179–200.
Tomasello, Michael. 2003. Constructing a language: a usage-based theory of language acquisition. Cambridge, MA: Harvard University Press.
Wasow, Thomas. 2002. Postverbal bahevior. Stanford, CA: CSLI Publications.
——. 2005. Intuitions in linguistic argumentation. Lingua 115(11).1481–96.
Whitsitt, Sam. 2005. A critique of the concept of semantic prosody. International Journal of Corpus Linguistics 10.283–305.
Wulff, Stefanie. 2008. Rethinking idiomaticity: a usage-based approach. London, New York: Continuum.
Wynne, Martin (ed.) 2005. Developing linguistic corpora: a guide to good practice. Oxford: Oxbow Books <http://ahds.ac.uk/linguistic-corpora/>.

Journals

Computational Linguistics: <http://www.mitpressjournals.org/loi/coli>
Computer Speech and Language:
<http://www.elsevier.com/wps/find/journaldescription.cws_home/622808/description>
Corpora: <http://www.eupjournals.com/journal/cor>
Corpus Linguistics and Linguistic Theory: <http://www.degruyter.com/journals/cllt/>
Empirical Language Research: <http://ejournals.org.uk/ELR/>
ICAME Journal: <http://icame.uib.no/journal.html>
International Journal of Corpus Linguistics:
<http://www.benjamins.com/cgi-bin/t_seriesview.cgi?series=IJCL>
Language Resources and Evaluation (formerly known as *Computers and the Humanities*):
<http://www.springer.com/linguistics/computational+linguistics/journal/10579>
Literary and Linguistic Computing: <http://llc.oxfordjournals.org/>