Methodological and interdisciplinary stance in Corpus Linguistics

Stefan Th. Gries, Professor of Linguistics at the University of California, Santa Barbara (United States), brings out a challenging notion of Corpus Linguistics. He proposes its understanding should be merged with psycholinguistic and cognitive concerns. Gries has no qualms in classifying Corpus Linguistics as a methodology. This explains his many references to methodological issues, ranging from the design of corpora to their comparison and/or analysis. In addition, Gries extensively discusses statistical issues, including how much knowledge a corpus analyst needs to have in order to embark on such an investigation. This sharp methodological concern is also expressed in his prospects for the practice in Corpus Linguistics, which, according to him, needs to develop from a statistical standpoint.

1. Where do you place the roots of Corpus Linguistics? And to what do you attribute the growth of interest in the area?

As to the first question, I am not sure the roots of Corpus Linguistics can be placed any particular place and/or time. As so often in science, related ideas emerge and develop in different places and then over time converge to give rise to a more coherent-seeming framework. It seems to me that the following are the most prominent early examples of what from today's perspective looks like corpus-linguistic work: bible concordances, Käding's (1897) work at the end of the 19th century, Firth's (1951) bearing on collocation, the Survey of English Usage as well as the Brown and LOB corpora, and all these are without doubt some extremely important milestones. Since it seems to me as corpus linguists are still more widespread or vocal in Europe, it may not come as a surprise that this list of highlights is very Euro-centric, so I would just like to add three American strands or approaches that I think should be included just as much.

First, there is the work of early American linguists. Not only did early Americanists such as Sapir rely on collections of utterances for their work, but so did American structuralists. For example, here is how Harris (1993: 27) describes Bloomfield's approach: "The approach […] began with a large collection of

recorded utterances from some language, a corpus. The corpus was subjected to a clear, stepwise, bottom-up strategy of analysis." Second, there is Charles C. Fries's compilation and analysis of a corpus to discover features of spoken American English (cf. Fries 1952), which was one of the first rigorously bottom-up, or corpus-driven, approach to the structure of (conversational) English. Finally, there is Zellig Harris's (1970:785f.) statement on distributional analysis which states more clearly than any other source I have ever seen the logic underlying most corpus- or computational-linguistic approaches involving co-occurrence data, i.e. concordances and collocations:

> [i]f we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution.

As to the second question, there is again not one single reason for the growth of interest. This growth has to do with several developments, again from different perspectives and at slightly different times. For example, there are logistic/structural reasons such that more and more corpora covering different languages, registers, etc. are becoming available, and the WWW is at our fingertips, so researchers can ask more and more diverse questions. Also, the field is maturing methodologically and conceptually: Corpus Linguistics was seen by many as consisting of little more than descriptive papers listing frequencies of occurrences of linguistic elements, but it is difficult for such onlookers to uphold that frame of mind. Not only do many corpus linguists use more and more sophisticated methods (for both retrieval and statistical analysis), but many corpus linguists are (finally …) beginning to look beyond the confines of the texts or discourses and contribute to, and interface with, neighboring fields such as cognitive linguistics/science, psycholinguistics, etc. Many of these fields also undergo a development towards more empirical/quantitative methods, which makes them compatible with corpus-based work. In a nutshell, they benefit from the data and methods we are dealing with all the time, and we benefit from them injecting a healthy dose of explanatory approaches and theoretical connections into our still too often merely descriptive discipline.

2. Is Corpus Linguistics a science or a methodology? Where would you situate Corpus Linguistics in the scientific or methodological panorama?

I have thought a lot about this question, especially since August 13, 2008. On that day, I received a first response to a call for enrollment for a quantitative Corpus Linguistics bootcamp that I was going to teach at my university, which triggered a

discussion that is now sometimes referred to as 'the bootcamp discourse' and that was, among other things, also concerned with that question. I think I advocate my position on this most clearly in my statement in the special bootcamp issue of the *International Journal of Corpus Linguistics* 15.3 (Gries 2010), but I will summarize it here very briefly while also taking up another related issue.

As for the first question, I think Corpus Linguistics is definitely a method(ology) or a 'methodological paradigm', no more but also no less. More specifically, to me *Corpus Linguistics* refers to (i) the study of the properties of corpora or (ii) the study of language on the basis of corpus data. I am making this difference here because I think as a corpus linguist, e.g. a corpus compiler, one can restrict one's attention to describing the frequency of linguistic phenomena in some corpora, the statistical properties of corpora, or even just methodological corpus issues (e.g., comparing how efficient different approaches to tagging a corpus are, or determining which kind of clustering algorithm best distinguishes different registers on the basis of *n*-gram frequencies) without necessarily being interested in a genuinely *linguistic* question, e.g., what the register differences actually reflect. Ultimately, I find the second type of study – addressing genuinely linguistic questions – more interesting, but investigations of the first type are still very important: Corpus Linguistics needs corpora and has not come up with many proven methods so compilation/sampling and methodological analyses are needed to prepare us for the second type of study.

I think the above distinction also bears on another way to situate Corpus Linguistics, namely with regard to the field of computational linguistics. While *computational linguistics* is only one of several terms to refer to a huge (and increasingly diverse) field – with *natural language processing* being one of the most widely used terms – some areas of computational linguistics of course border on, or overlap with, Corpus Linguistics. When asked where I see the (main) difference between these fields, I usually say (a bit polemically and simplistically) that some areas of computational linguistics are in fact mislabeled: taking the notion of head-modifier structure very literally, I think there are many areas that should be labeled *linguistic computing* as opposed to computational linguistics, and this distinction also relates to Corpus Linguistics. My take on this is that I want to call something ___ *linguistics*, if its ultimate goal is to increase our understanding of (the use of) human language, or even the linguistic system's place in the larger domain of human cognition, and I want to call something ___ *computing* if its ultimate goal is not concerned with *understanding* (the use of) human language but its computational application or implementation. For example, for me, developing a talking ticketing machine for the airport parking lot falls under the heading of *natural language processing*, but I would not call it ___ *linguistics* (even if frequency data from corpora are used to tweaks how the machine parses its input),

but, if pressed, would call it *linguistic computing*. (Of course, there are cases where such a forced binary decision is difficult to make.)

As for the second question, as I have argued in Gries (2010), Corpus Linguistics should be "a psycholinguistically informed, (cognitively-inspired) usage-based linguistics which should be located, firmly and deliberately, in the social/behavioral sciences." As mentioned above, it is time for more of the field to move beyond the purely descriptive and sometimes ostrich-like we-must-not-look-beyond-the-texts approach and assume (proudly, I might add) the position that our discipline deserves: we are looking at complex behavioral data typically arising from social settings, which means we should describe what the data look like with decent quantitative methods and explain their nature with reference to findings from relevant fields, and fields that are obviously relevant for a discipline studying *behavioral* data from *social settings* are cognitive science/linguistics, psychology/-linguistics, and sociology/-linguistics plus their respective neighboring fields.

3. How representative can a corpus be?

The answer to this question obviously depends on one's definition of *representative* so let me first clarify how I understand *representative*; not everybody uses it to refer to the same thing. I would call a corpus representative if it contains samples of all the different parts of the linguistic population that the corpus is supposed to represent. I would call a corpus also balanced if the sizes of the samples of the linguistic population it contains are proportional to the proportions these parts make up in the population that is supposedly represented in the corpus.

With these definitions in mind, my assessment is rather pessimistic: I think a corpus can be somewhat representative *on some level* largely by virtue of its design, but balanced probably only largely by virtue of sampling luck, but even this statement needs to be qualified some more for two reasons. First, because I think the degree to which a corpus can be representative and balanced is correlated with its position on the general-special continuum of corpora: given a particular amount of resources, the more specific the corpus is intended to be, the more representative and balanced it can be; and the more general it is supposed to be, the less representative and balanced it will be.

Second and more importantly, this statement needs to be qualified because, as with all corpus work, there are innumerable nested levels of granularity that can be considered. Theoretically, sampling for corpus compilation is a multidimensional enterprise even though, for obvious and reasonable practical purposes, only a small number of dimensions can be chosen. For instance, as I understand it, corpus compilers usually (and reasonably so) select a sampling

scheme whose units involve modes (spoken vs. written or a more fine-grained version of this) and, within the modes, registers. The design of the ICE-GB, for example, involves three different levels of hierarchical (sampling) organization, as indicated in the columns of Table 1 (cf. <http://www.ucl.ac.uk/english-usage/projects/ice-gb/design.htm>).

Table 1. The hierarchical organization of the ICE-GB

Mode	Register	Sub-register
Spoken	Dialog	Private vs. public
	Monolog	Scripted vs. non-scripted
	Mixed	Broadcast
Written	Printed	Academic vs. creative vs. instructional vs. non-academic vs. persuasive vs. reportage
	Non-printed	Letters vs. non-professional

All of this raises three problems that can decrease the representativity of a given corpus and, hence, its balancedness. A corpus will be representative and balanced to the degree that the corpus compilers succeed in

- identifying the relevant levels of corpus organization, i.e., the adequate columns in Table 1 and their right number;
- identifying the relevant distinctions within each column of corpus organization, i.e., the distinctions indicated in the column-specific rows of Table 1;
- determining the sizes of each of the samples that result from the different levels and their within-level distinctions.

As for the first two problems, even if corpus compilers managed to identify all the right registers – i.e., made the corpus representative on the level of register –, this does not guarantee that the way they would sample the registers does not make the corpus *un*representative on the level of the sub-register and/or, even worse, on many other levels. Strictly speaking, even if corpus compilers succeeded in choosing the right modes, registers, and sub-registers, then it is strictly speaking still possible that the way they sample from texts on the level of the sub-register was unrepresentative. For instance, if one chose only the first and the last sentence of each text/conversation or if the texts from which one sampled exhibited untypically large sentence lengths or a near complete absence of a particular construction *C*, then the corpus would be representative down to the level of the sub-register, but unrepresentative with regard to discourse features, sentence lengths, or the frequency of *C*. (Of course, corpus compilers would not just choose the first and/or last sentence, this is just a hypothetical example: any one linguistic variable could be used as an example.)

As for the third problem, even if the corpus compilers managed to make the corpus representative on many levels, they could still make it very unbalanced because they might not succeed in getting the sample sizes right. And how would they get them right anyway, how do we determine the proportional sizes of the samples – in terms of speaking time, in terms of utterances, sentences, words?

In sum, I think it is possible to achieve some degree of representativeness and balancedness when compiling a general corpus, but only on some level(s) of corpus granularity. A corpus that is perfectly representative and balanced on one level can be completely unrepresentative in terms of the frequency distribution of some specific pattern. Strictly speaking, it is therefore necessary to sample as widely as possible and explore for each phenomenon of interest how it is distributed over multiple levels of corpus divisions especially since the most meaningful division of a corpus into parts may be different for each phenomenon and may not coincide with linguists' favored register distinctions.

4. How far should an analyst rely on intuition?

My take on this is that intuition can play a role on nearly all levels of corpus-linguistic analysis, and it often has to, but of course to varying degrees, and the following comments adopt a broad notion of intuition (itself a fuzzy word), one that involves all sorts of subjective decisions. In general, there is a subjective decision that is sometimes overlooked when the subjectivity of an analysis is evaluated, and that is which corpus or which genre/register to study. In addition, intuitive/subjective decisions come into play at different points of time.

First, the identification of a topic or problem typically involves a lot of intuition such as when a researcher finds that the explanation of phenomenon *P* does not appear satisfactory given what else is known about *P*. A little less intuition would be involved when a researcher finds that the explanation of *P* is unsatisfactory given a new set of data.

Second, the retrieval of data may involve very little intuition as when no decision for a particular corpus has to be made (because, say, only one is available) and one looks for a uniquely identifiable word form. More intuition is needed when a decision for some corpus (or other database) has to be made, but also when one looks for partially lexically-filled constructions such as the *into*-causative (NP_{SUBJ} V NP_{DO} *into* V-*ing* as in *He tricked her into marrying him*) where one might decide to search a corpus for '`\binto [^\s]+in['g]\b`'[1] and then use the linguist's 'intuition' (often called *knowledge*) to weed out false hits such as *This is how the*

1. This regular expression matches a word boundary ("\b"), followed by *into* ('into'), a space (' '), some characters that are not spaces ('[^\s]+'), *in'* or *ing* ('in['g]'), and a word boundary ('\b').

EU came into being. And even more intuition is needed when the linguistic phenomenon of interest does not involve anything literal to search for, i.e. no specific word and/or tag or involves an unannotated corpus.

Third and most importantly, subjective decisions will become necessary during the analysis, i.e. during, for example, the coding of data with regard to features that are not always clear-cut, and during the statistical analysis of the data. As for the former, if one wants to code lengths of utterances, one must choose between counting characters, morphemes, words, phrases, etc. If one wants to code referents of noun phrases (NPs) semantically, one may have to distinguish between concrete and abstract, but within the former one can again distinguish animate and inanimate. In this case, do humans get a category on their own? Do animals, or plants? What do we do with NPs referring to actions (as in *Counting is hard*)? Do we use Vendlerian categories for that (which are in turn difficult)? In the sentence *the police came to the crime scene*, is the subject concrete/human or abstract/organization? And what about *minefield* in *wading through the minefield of autism treatments* – is it locative, or do we code the whole thing as an idiom? Maybe just as difficult is the coding of coherence relations, or the coding of referential distances (do we include cover terms or not, part-whole relations or not, etc.). It is these kinds of tricky decisions that have resulted in more and more studies including inter-rater reliability statistics in their papers (not that these are completely unproblematic, but that is a different story).

As for the statistical analysis, one sometimes also has to make decisions regarding the method to be adopted. For example, which similarity measure and which amalgamation rule to use in a cluster analysis? Or what kind of cluster analysis to use in the first place: hierarchical or phylogenetic? For example, which method to use to predict an alternation – logistic regression, classification trees, Bayesian classifiers etc.?

In sum, it is obvious that corpus linguists need to make subjective decisions all the time, and they need to document their subjective choices very clearly in their publications. However, in spite of these undoubtedly subjective decisions, many advantages over armchair linguistics remain: the data points that are coded are not made-up, their frequency distributions are based on natural data, and these data points force us to include inconvenient or highly unlikely examples that armchair linguists may 'overlook'.

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

I am not sure I am in a position to tell any researcher what one *should* think of as one's primary research question, so I will instead mention two questions one should think of which concern and/or qualify the scope of one's primary research question and findings.

The first of these questions has to do with the kinds of results reported and what the sources of their variation in corpus data are. More specifically, given that corpora only provide frequencies of (co-)occurrence, studies usually provide (conditional) frequencies, means, and averages of the phenomenon in question. However, obviously each corpus and each part of a corpus will yield different results, and these differences will sometimes reflect something linguistically interesting, and they will always also reflect random variation of sampling. I would therefore like to see more exploration of how variable and sensitive such results are. Schlüter (2006) compared the widely differing frequencies of present perfects in very different corpora, motivated me to explore how widely the frequencies of present perfects differ within (parts of) *one* corpus. The results of that exploration – the frequencies of present perfects in the ICE-GB alone were just about as variable as those in very different corpora – plus some follow-up work (in Gries 2006) provided some sort of an epiphany for me: claims about an overall frequency of (co-)occurrence, a mean, or a correlation can be useless unless they are accompanied by an exploration of the diversity of the data giving rise to the overall frequency, the mean, or the correlation. It is this kind of systematic and often bottom-up exploration of different levels of granularity that I think is essential for our understanding and validation of virtually all corpus results.

The second of these questions has to do with what corpus-based results are used for, and with the question of what Corpus Linguistics is. As mentioned above, studies that, for instance, describe the frequencies of some linguistic element(s) in one or more corpora fall under my heading of Corpus Linguistics, but I also said that I prefer studies that try to go beyond that. Put differently, I prefer studies that describe something but then also answer 'why does that happen?' questions, which try to understand the motivations and the forces driving the distributions of data, and which ideally try to do this by exploring connections to findings from *other fields* to avoid the often circular description (often not even reasoning) that arises from some gatekeepers' reluctance to admit other kinds of evidence onto the corpus linguist's desk. It may be in this regard that I disagree most strongly with some other scholars' beliefs. For example, when Teubert (2005:10) writes "When linguists come across a sentence such as 'The sweetness of this lemon is sublime', their task is […] to look to see if other testimony in the discourse does or does not provide supporting evidence," I cannot even begin to understand why that should be the task of any corpus linguist: what would be explained by this? Similarly, I cannot subscribe to his statement that "[c]orpus linguistics […] is not concerned with the psychological aspects of language" (Teubert 2005: 2f.). Although there are limits to what corpus linguists can say about the human mind and its psychology, that does not mean that distributional data from corpora cannot inform, or be meaningfully related to, data from more psychological/cognitive disciplines.

For example, is it not better to be able to explain distributions in corpora – of, e.g., reduced pronunciations of words – with reference to generally-known cognitive mechanisms regarding learning, habituation, and articulatory routines than to point to other things happening in the discourse? Is it not interesting to be able to explain changes in diachronic corpora – e.g., the development of *going to* as a future marker in English – with reference to generally-known effects of automation as a result of frequency?

In sum, I do not really dare make specific research recommendations, but I would love to see corpus linguists be more aware of, and explore in a bottom-up fashion, the variability of the data they report on as well as establish more explicit (and explanatory) connections of their descriptive results to findings from other disciplines.

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

To my mind, the two most important advantages of corpus work are the following. First, the data come from authentic settings: conversations and texts that were produced in largely natural contexts. While that makes corpus data very messy and noisy compared to experimental data (which then of course are in turn potentially more tainted by the artificiality of the experimental setting), it also enriches them and allows us to include cotextual and contextual/situational aspects of language use in our analyses.

Second, corpora only provide statistical data – even if no proper statistical analysis is conducted – and that means that statements such as 'in corpus *C*, 8.5% of X were Y, compared to 22.1% of Z' can be straightforwardly tested for replicability, compared to other corpus or experimental studies, be extended by additional data, and tested for significance, whereas armchair statements of the types 'X is rather untypical' or 'X is marginally acceptable' fare much worse in these respects.

In terms of weaknesses, or maybe risks, of corpus analysis, I see a few of those, but many of them are not peculiar to corpus analysis but apply to many empirical settings. For instance, one must bear in mind that whatever findings one reports that one can only generalize from the studied sample to a larger population to the extent that the corpus is representative with regard to the targeted population. Unfortunately, there are some authors who are quite happy to generalize more liberally.

Second and in a related manner, while corpus data are usually samples from naturally-produced texts as mentioned above, one needs to be aware of the fact that the circumstances of these texts can still be at odds with one's research question. For example, given the easy availability of large amounts of journalese data, many corpus studies use them, and often this is a good thing. However, as we

argued elsewhere (cf. Gilquin & Gries 2009), even if corpora consisting of journalese data only may be large, they are still rather unsuited as a general corpus since they are a very peculiar register: they are created much more deliberately and consciously than many other texts, they often come with linguistically arbitrary restrictions regarding, say, word or character lengths, they are often not written by a single person, they may be heavily edited by editors and typesetters for reasons that again may or may not be linguistically motivated, etc. Thus, the more such characteristics can undermine one's research purpose, the more one must hedge the generalizability of one's findings or turn to additional (corpus or experimental) data for validation.

Finally, on the most fine-grained level of specific analyses, I am sometimes rather unhappy with the methodological decisions made by some analysts. On the one hand, the level of statistical sophistication of quite a few studies leaves much to be desired, with the two most pressing issues being (i) the complete lack of statistical significance testing (ignoring for now the problems that may come with significance testing) and (ii) the problem that multifactorial phenomena are studied monofactorially, disregarding the nature of often complex interactions of factors. On the other hand, there are (thankfully fewer and fewer) studies that can never be characterized better than in Pullum's (1978: 400) words: "The fault is the procedure of attempting to establish a case on the basis of a set of data the size of a small workbook problem (though with theoretical biases of more generous proportions)." I have seen many papers which made far-reaching claims regarding a frequent phenomenon/word based on perhaps 200 examples. When I explain to my students why I hate that kind of practice, I tell them, "How come corpus linguist X thought, 'Gee, let me look at 150 examples this afternoon, surely that will be enough …' while biologists try to grow some cultures for months, archaeologists try to dig up stuff for years, etc. We look at, code/annotate, and evaluate examples and their frequencies, so how come some assume looking at 150 examples on one afternoon does the job?"

Thankfully, my perception is that the field is maturing more and more and addressing these shortcomings in various ways. Still, all of us, me included, still have a long way to go…

7. What is the future of Corpus Linguistics?

I do not think I have a good answer to that question. There are some currently hot topics but I think it is pretty much impossible to even make a reasonably precise educated guess, given that scientific disciplines do not exactly evolve nicely linearly. I will therefore only offer some brief 'guesstimates'.

In terms of areas/topics, I think corpus-based research will play an increasing role in applied linguistics, especially with the growing number of learner corpora and the ever increasing interest in second language acquisition and teaching. Similarly, I expect to see a greater degree of convergence between Corpus Linguistics and sociolinguistics, given how these disciplines share commitments to authentic data and quantitative analysis. Also, there seems to be a growing interest in corpus-based methods in the fields of language description and language documentation, which involves long-term digital archiving to store data on often endangered languages. Obviously, this ultimately raises issues of formats and annotation, but at the same time this increasing availability of such data will doubtlessly stimulate more desire to retrieve data from such databases or corpora for linguistic analysis, and I would hope that both fields can help each other evolve. On the one hand, corpus linguists have long thought about matters of corpus formats, storage, annotation, and access and have learned many lessons – especially from corpora much larger than those handled in language documentation – that documentary linguists could benefit from. On the other hand, documentary linguists routinely deal with languages whose structural complexity poses complicated but interesting annotation challenges that corpus linguists, who have mostly (but not exclusively) worked on the usual suspects from the Indo-European language family with often much more impoverished morphologies.

Second, I think that the current trend of using corpus data in psycholinguistic and cognitive approaches will become stronger. As for the former, language acquisition research has long involved corpus data, but there is also more and more work on probabilistic approaches to language production and comprehension, and much of this work is based on frequencies of words, *n*-grams, and constructions from corpora. The number of articles in the *Journal of Memory and Language* that mention corpora has risen considerably over the past few years. This trend can also be seen in cognitive-linguistic approaches. That whole field is taking the notion of usage-based approaches more and more seriously,[2] and the number of submissions to *Cognitive Linguistics* that involve corpus data has been on the rise. Since there are now also more corpus linguists talking about such issues and seeking explanations that transcend the narrow boundaries of pure corpus description, I would hope that this marks the beginning not of a convergence of these fields, but of recognition of what these fields have to offer each other as well as more fruitful mutual collaboration.

2. The term *usage-based* is used in the sense of assuming that human linguistic systems are affected by the actual use of linguistic elements and structures, which makes for an obvious connection to corpus-based work.

Finally, Corpus Linguistics will mature statistically. I know of maybe one multifactorial corpus-based study of a syntactic alternation from before 2000 (Gries 1999, later improved on in Gries 2003), but now binary logistic regressions, mixed-effects models, cluster analyses, etc. are not uncommon anymore and can be found in nearly every corpus-linguistic journal. Doubtlessly, and fortunately, this trend will continue.

8. What is the role of programming knowledge in undertaking corpus work?

Unfortunately, the role is rather limited, and that, together with the absence of proper statistical training, is the largest methodological problem of this discipline. Just look at the situation from an unbiased observer's perspective: why is it that corpus linguists often must retrieve complex patterns from gigabytes of messy data in various languages, encodings, forms of organization and with widely differing forms of annotation, but most curricula do not contain even a single course on basic programming skills or relational databases (while psychologists, computational linguists, cognitive scientists etc. devote years to acquiring the required methodological skills)? It is true that there are several tools that allow users to perform a few elementary corpus-linguistic tasks with a graphical user interface but, while I am not evaluating these programs here, let me say this quite bluntly: the superficial richness of functions and buttons is deceiving and debilitating. I know colleagues whose corpus-linguistic skills are defined by what WordSmith Tools (or AntConc, or ConcGram, etc.) or, even worse, web interfaces can do – if you take whatever resource they use away from them, they cannot pursue their corpus studies anymore. This means that if these corpus linguists' program(s) cannot lemmatize or use regular expressions, or compute keywords for *n*-grams (with $n \geq 2$), neither can they. If their program does not allow them to conflate several files into one or compute particular collocational statistics, neither can they, etc. Does anyone know any other scientific discipline where this is the case, where quite a few practitioners' main methodology is opening a (corpus) website in their browser?

Thus, the field must step it up a bit and go beyond commercial software (and websites) because, first, the software many people use is severely limited in terms of

– availability: not every program runs on every operating system and not every researcher can afford the program(s) they would want to use;
– functionality: programs/websites can only do what is hardwired into them so if the program/website cannot compute collocational statistics, handle Unicode, corpora with particular tag formats, corpora with stand-off annotation

or multi-tier annotation, then a (commercial) program or website, rather than our interests/needs, dictate our research agenda!

– user-control: users are at the mercy of the developers. If, for instance, the creator of a program updated the way key words are computed without informing the users, then users would be clueless as to why the same data set suddenly yields different results, and with non-open source software, one cannot find out what has happened and why. Or, if the developer of 'your' program decided to discontinue its development or a Microsoft Windows update changes a .dll so the program stops working, then what? Or, finally, to return to the previous point: how are we even going to make any progress in the field? How would one study whether minimum sensitivity or $p_{\text{Fisher-Yates exact test}}$ or ΔP are better collocational statistics than *MI* and the loglikelihood statistic if one is utterly dependent on one tool which happens to not offer these measures? And are we happy with the fact that this situation would put the vast majority of the field under the control of two or three people who happen to develop nice-looking software? MonoConc Pro's current version – 2.2 – has been around with this version number since 2002 but I for myself am glad that my methodological knowledge has advanced a bit since then…

Second and again quite bluntly, inflexible software creates inflexible researchers: more methodological knowledge sometimes suddenly suggests ways of analysis one would not think of, given how one's dependence on a ready-made tool can restrict one's way of thinking about a problem. Put differently, with a programming language, one does not need to think outside the box – because there is no box: everything's possible. Recently, I was involved in a project where we needed to recover sequences of two or more adjectives in a learner corpus. However, the corpus was not tagged, which for many colleagues would mean they would not be able to do the study. In our project, we used a small R script that searched the whole BNC for *all* words tagged as adjectives, saved them into a list, and then added an adjective tag to every occurrence of a word from that list in the learner corpus. Thus, we could then simply search for sequences of two words tagged as adjectives. A maybe even more telling example involves the search for ditransitive constructions, again in an untagged/unparsed learner corpus. As a heuristic, we used a script (less then 60 lines) that recovered all verb tokens tagged as used ditransitively in the ICE-GB, looked up the lemmas for these tokens in a lemma list, looked up all the forms for these lemmas in the lemma list (to get *allocating* as a search term even if only *allocated* had been used ditransitively in the ICE-GB), and then outputted a concordance of all matches of those forms in the learner corpus. This is not perfect, but it is easy to see that no ready-made program could

ever do this (especially not quickly). Thus, it is absolutely imperative for the field to further evolve in this direction, and fast, please![3]

9. What are the issues involved in comparing corpora?

In some sense, I have already touched upon this issue in some of the above questions and, as I will argue below, I think the question should actually be asked slightly differently. By way of preface, this topic's importance is so large as to only be matched by the degree that it is understudied. The reason for why this topic is so important is twofold. First and as mentioned above, for any phenomenon every corpus and every part of a corpus will yield different results, and whatever results we report, they will come with some degree of variability. Second and as a consequence of that, we need to assess the variability of the results we obtain against differences between corpora, between parts of different corpora or of one corpus (i.e., corpus homogeneity), and between corpora and the linguistic population they are supposed to represent (i.e., we again face the problems of representativity and balancedness), which are all inextricably related. Thus, questions regarding "the issues involved in comparing corpora" should actually *always* be phrased as regarding "the issues involved in comparing (parts of) corpora."

I said the topic is unbelievably under-researched, and this is so for three issues, which the present question is concerned with. The first issue is the complexity arising from the interrelations of these various kinds of differences. If a corpus

3. One reviewer suggested a parallel between a linguist's computer program and a doctor's instrument to ask whether the former would need to develop a tool of his/her own. This is wrong on so many levels that I hardly know where to start. Some of the above points should already illustrate why, but here is a different take on this. Many doctors practice medicine work in an applied field, where they use a set of finite heuristics to quickly identify a patient's illness, often out of a small and finite set of possibilities. They typically use several instruments (e.g., a stethoscope), which are often highly specialized (performing just one function), which do not differ much as to how well they perform that function, and which do not undergo developments and updates that do not allow to replicate results. If a stethoscope does not yield the desired result, the doctor uses another tool from the large set of available tools. By contrast, many corpus linguists work in a fundamental research field, where they must develop a research strategy in many different steps to describe/explain a phenomenon, usually involving an open-ended set of alternatives. Thus, the above-mentioned methodological and conceptual imprisonment caused by the few non-customizable functions of commercial software impedes the development of the research strategy *and* the procedure of arriving at, and interpreting, the results. And if the commercial software is not designed to produce the desired results, then the corpus linguist without programming experience either has to live with a potentially foul compromise or drop the project? This is not acceptable: every corpus-linguistic researcher should have some programming skills.

yields an overall result then this result may only be really worth considering at the level of the corpus if

– the corpus parts did not yield completely different results such that the one on the level of the corpus is only an unrepresentative average of very different results from corpus parts;
– the corpus is representative (and maybe balanced) enough with regard to the language or variety or register it is supposed to represent so that we may assume the corpus result will speak to what happens in the population.

The moral therefore is to bear in mind, in corpus compilation and analysis, that there are many different levels of corpus granularity: varieties, registers, files, texts, and within- and between-corpus comparison should take all of them into consideration.

The second issue is the fact that we still do not know yet which statistics are best suited for the comparison of corpora. There are studies that have begun to address this notion by proposing, reviewing, and/or exploring a variety of statistics that could be used; other studies approach the issue with different simulation/resampling-based approaches, but this problem is far from resolved (cf., e.g., Kilgarriff 2001, 2005, who argues against significance testing, and Gries 2005, who demonstrates that some of Kilgarriff's objections are mistaken). Thus, we need more exploration of statistical methods for corpus comparison, but also – a very general problem of Corpus Linguistics – much more validation of new *and* existing methods.

The final issue is the fact that, with very few exceptions, the little work that is out there only addresses a single level of corpus granularity and corpus comparison: the word. This has to do with a general bias of corpus linguists to study words, or lexical items, and it has to do with ease of retrievability of these elements (especially in the usual suspects of Indo-European languages that most corpus linguists work with, where words can be identified more easily than in polysynthetic and fusional languages). However, since (parts of) corpora can differ on any level of linguistic granularity and, somewhat ironically, it is corpus linguists and cognitive linguists who now assume that words are not different in kind from more schematic patterns/constructions, corpora that seem very similar on the level of the word may be very different on the level of other linguistic expressions. Thus, corpus comparison has to not only take differences arising from the granularity of the corpus/corpora and its/their parts into consideration (cf. issue #1 above), but also differences arising from the (level of) linguistic phenomena whose frequencies are used for comparing corpora and/or their parts. Thus, this answer is again a plea for more systematic bottom-up exploration of where similarities reside and what their implications are.

10. How much statistics does a corpus analyst need to master?

In a sense, there are two answers to this question. Superficially, the first answer is that this of course depends on what exactly a corpus linguist's focus is on. It would seem that a corpus linguist studying a word *w* by means of a highly qualitative analysis of the contexts of *w* in a small sample of newspaper texts (e.g., two texts each from ten consecutive years) does not need statistical expertise. However, as I have argued frequently, this view is fundamentally mistaken. Again, just look at the situation from an unbiased observer's perspective: why is it that corpus linguists look at something (language) that is completely based on distributional and probabilistic data and just as complex as what psychologists, psycholinguists, cognitive scientists, sociologists, etc. look at, but most of our curricula do not contain even a single course on statistical methods (while psychologists etc. regularly have two to three basic and one or two advanced courses on such methods)?[4] To understand more concretely why this is a huge problem, let's assume a linguist finds that *w* appears to be used increasingly negatively over time. The questions that immediately arise from this assumption are (i) how did the linguist find that and (ii) how can or must this be interpreted.

As to the first question, it is important to realize that corpora do not provide such findings without recourse to frequencies because corpora provide nothing but frequencies of (co-)occurrence and dispersions. Thus, a corpus linguist can only infer that *w* is used increasingly negatively because the percentage of times *w* is used with negative collocates out of all uses of *w* is becoming larger over time. Thus, whatever pattern or function, meaning, or use is inferred from corpora, it is based on distributional data, and the science that tells us how to handle distributional data best is statistics. But while percentages are of course not exactly a most sophisticated statistic requiring expertise and training, there is still the second question, which is concerned with how the increase in relative frequency is interpreted.

4. When asked about how future generations could learn statistics if it is not part of the curriculum, I have two answers to this question: one is serious, the other is serious, too, but also has a 'duh' attached to it. As to the former, the number of researchers who have realized how important statistical training is increasing and so is the number of places where statistical training is offered. Also, there are now several venues – workshops, bootcamps, etc. – in which researchers and students can begin to take their first steps under supervision, just as they can spend time as a visiting scholar/student in departments where such resources/people are available. As for the latter answer: when there are no such opportunities – which was actually the case for my own linguistic upbringing: at the time, the department from which I obtained my degrees offered no Corpus Linguistics or statistics training at all – then there is still another approach, which is somewhat old-fashioned but has worked well for me: it is called 'self-study' (books and other resources)… Where, if not in academia, is living to be equated with learning?

As to the second question, two points must be considered. First, is the trend statistically significant, i.e., whether pronounced enough to probably not just result from random variation in data? To answer that question, one needs to decide on whether to use a correlation coefficient (and if so, which one to use), or whether to group the data into, say, two, four, or five groups of ten, five, or four percentages each and do a comparison of means (and if so, which test for means to use) or a comparison of observed frequencies (maybe with a chi-square test). Deciding which correlation coefficient to use (e.g., Pearson's r, Kendall's τ, etc.) or which test for means to use (e.g., a t-test, a U-test, a one-way ANOVA, etc.) or whether a chi-square test can be used in turn requires knowledge of notions such as normality, variance homogeneity, maybe the central limit theorem etc. so it is not clear how even something as simple as distinguishing a change of semantic connotations over time from random variation can be done without statistical knowledge. (And I do not even mention issues such as independence of data points etc. here…)

Second, let's assume the computation of Kendall's τ shows there is a significant upwards trend. This still leaves open the possibly interesting questions what kind of trend (linear or nonlinear?) and whether that trend is specific to *w* or whether *w*'s collocational behavior is just one reflection of a more general trend. What if *w* is the word *Muslim*, but in the wake of 9/11, religiously-motivated conflicts throughout the world, and the recent financial and abuse scandals of the Catholic Church, words from the semantic field of religion, or even of any larger organization or group, are reported on more negatively? Would one want to make a claim about how newspaper coverage on Muslims has become more negative over time when in fact newspaper coverage on *all* religions has become more negative? I do not think so, and thus one needs one or more additional samples of collocate frequencies on words referring to, say, a Christian religion and some other religion and then do a statistical test to see whether *Muslim* is special and worthy of much individual discussion in this context or whether *Muslim* is just one example of a general trend. Again, this cannot be done without statistical knowledge (about, here, regressions or linear models), and it has always completely escaped me how there are still people who cannot see this…

References

Firth, J. R. 1951. *Papers in Linguistics*, 1934–1951. Oxford: OUP.

Fries, C. C. 1952. *The Structure of English: An Introduction to the Construction of English Sentences*. New York NY: Harcourt Brace.

Gilquin, G. & Gries, St. Th. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory* 5(1): 1–26.
Gries, St. Th. 1999. Particle movement: A cognitive and functional approach. *Cognitive Linguistics* 10(2): 105–145.
Gries, St. Th. 2003. *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London: Continuum Press.
Gries, St. Th. 2005. Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory* 1(2): 277–294.
Gries, St. Th. 2006. Exploring variability within and between corpora: some methodological considerations. *Corpora* 1(2): 109–151.
Gries, St. Th. 2010. Corpus Linguistics and theoretical linguistics: A love-hate relationship? Not necessarily… *International Journal of Corpus Linguistics* 15(3): 327–342.
Harris, R. A. 1993. *The Linguistics Wars*. Oxford: OUP.
Harris, Z. S. 1970. *Papers in Structural and Transformational Linguistics*. Dordrecht: Reidel.
Käding, F. W. 1897. *Häufigkeitswörterbuch der deutschen Sprache*. Steglitz: no publ.
Kilgarriff, A. 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6(1): 1–37.
Kilgarriff, A. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2): 263–276.
Pullum, G. K. 1978. Assessing linguistic arguments by Jessica R. Wirth. *Language* 54(2): 399–402.
Schlüter, N. 2006. How reliable are the results? Comparing corpus-based studies of the present perfect. *Zeitschrift für Anglistik und Amerikanistik* 54(2): 135–148.
Teubert, W. 2005. My version of Corpus Linguistics. *International Journal of Corpus Linguistics* 10(1): 1–13.