Balancing Acts: Empirical Pursuits in Cognitive Linguistics

John Newman

Abstract

Cognitive linguists, more so than linguists of other persuasion, are expected to be open to a wide range of data and methodologies. Usage-based data and methodologies are playing an increasingly important role in cognitive linguistics and it is appropriate to re-examine some of the common practices in dealing with this kind of data. It is argued that a number of practices in the use of corpora and associated methodologies should be evaluated more critically than they currently are. This does not mean that we must reject current practices, but we should be more aware of the need to weigh alternatives more carefully before we settle on one way of proceeding.

Keywords: usage, corpus, spoken corpus, conversation, communication, part of speech tags, quantitative methods, inflected form, lemma

1. Introduction

The terms *cognitive linguistics* and *cognitive semantics* are open to a variety of interpretations, but here I will be guided by the characterizations given in Evans and Green (2006). Broadly speaking, cognitive linguistics is concerned with general principles that provide some explanation for all aspects of language, including principles drawn from disciplines other than linguistics (Evans and Green 2006: 27-28). If we accept this characterization of the field, then many kinds of evidence and methodologies are relevant. Clearly, many "balancing acts" are required if we are to do proper justice to all the data which is potentially relevant to a cognitive linguist, so understood. Cognitive semantics is a subfield of cognitive linguistics, but is hardly less general, concerned as it is with "investigating the relationship between experience, the conceptual system and the semantic structure encoded by language" (Evans and Green 2006: 48). The three phenomena referred to in this characterization of cognitive semantics–experience, the conceptual system, semantic structure encoded by language–also suggest a variety of techniques and methodologies for their full elaboration. In other words, whether one is pursuing cognitive linguistics or a more narrowly defined cognitive semantics, there will have to be some balancing acts required on the part of the researcher in order to reconcile the diversity of evidence.

Rather than attempt an overview of all the issues we face as we try to navigate our way through all the possible types of data and methodologies which cognitive linguists or cognitive semanticists might avail themselves of, I will focus on one kind of data and the methodologies associated with it, namely corpus data. Here, corpus data is understood as the collection of relatively large amounts of connected (transcribed) speech or writing. In line with the broad definition of cognitive linguistics above, this kind of data and its associated methodologies are allowed for in the practice of cognitive linguistics and cognitive semantics. Indeed, Evans and Green (2005: 108) draw attention to the importance of usage-based data in cognitive linguistics, a point taken up further in Section 2 below: "language structure cannot be studied without taking into account the nature of language use". Geeraerts (2006: 29) speaks of the "growing tendency of Cognitive Linguistics to stress its essential nature as *a usage-based linguistics*". In light of this particular claim about cognitive linguistics, it follows that corpus data drawn from actual usage is not only permissible, but necessary in cognitive linguistic and cognitive semantic research.[1] This is a point worth clarifying at the outset, since one could imagine that a field called "cognitive linguistics" or "cognitive semantics" would not possibly be concerned with the external products

of usage. Teubert (2005: 8), for example, seems to understand "cognitive linguistics" in this way and, not surprisingly, considers corpus linguistics and cognitive linguistics to be "two complementary, but ultimately irreconcilable paradigms". The reality of practice in the field known as cognitive linguistics, however, is, for better or for worse, the reality captured by the characterization quoted above from Evans and Green (2006), whereby every attempt is made to explore and reconcile a range of evidence from quite varied sources (cf. Tummers, Heylen, and Geeraerts 2005).

Evans and Green (2006: 153-467) provides a full overview of the issues which are of interest to cognitive semanticists. It would be impossible to review the use of corpus techniques as they might apply to all these issues. Instead, I will address a number of quite fundamental issues which are relevant to the use of corpora in linguistics generally, but particularly with respect to cognitive linguistics. In what follows, I question a number of current issues: over-reliance on language data abstracted from the context of communication (Section 2), methodological issues involving registers (Section 3), inflected forms vs. lemmas (Section 4), reliance on part of speech tags (Section 5), and use of statistical measures (Section 6). My intention is to draw attention to alternative ways of carrying out corpus-based studies, alternatives which linguists of a cognitive linguistic persuasion, as understood by Evans and Green, should be open to. While the focus of this volume is cognitive semantics, I will address issues which, I believe, have relevance for the larger field of cognitive linguistics, not just for cognitive semantics.

2. The communicative context

One issue which deserves more discussion than it receives in cognitive linguistics, and more generally linguistics, is the scope of our inquiry. Despite all the subfields of linguistics that exist, the field as a whole remains preoccupied with the study of abstract structure at the expense of the study of communication. Indeed, the study of abstract linguistic structure has come to virtually define the discipline of linguistics. This imbalance will arguably always remain a feature of the field, but it does not follow that the study of communicative activity should be ignored to quite the extent that it is. Cognitive linguists have been more open to incorporating contextual information in their analyses than linguists from most other schools of thought. Even with this openness to a more inclusive view about language and context, however, the study of the entire communicative contet seems to elude cognitive linguists.

The "usage-based hypothesis" already alluded to is sometimes held to be a key idea guiding research in cognitive linguistics and one might expect that hypothesis to lead to the study of communication at a primary concern. Evans and Green (2006) elevate the usage-based hypothesis to one of two fundamental assumptions in cognitive linguistics:

(1) a. The first guiding assumption [of the cognitive approach to grammar] is the symbolic thesis which holds that the fundamental unit of grammar is a form-meaning pairing or symbolic unit. (p. 476)

b. The second fundamental assumption of the cognitive approach to grammar is the usage-based thesis... [this thesis] holds that the mental grammar of the speaker ...is formed by the abstraction of symbolic units from situated instances of languages use. (p. 478)

Evans and Green (2006) point out that their usage-based hypothesis leads to a rejection of any strict division between "competence" and "performance", or knowledge of language and use of language. While this view would certainly allow for a broadening of the scope of inquiry, usage as such has not been a particularly common research focus in the field as a whole. In some

other publications which purport to give an overview of cognitive linguistics, usage does not occupy any privileged place. Even in its second edition, Dirven and Verspoor's (2004) introduction to the field has no extended discussion of the concept and the term *usage* is not listed in the index. The closest they come to a discussion of usage is a review of Gricean maxims and Searle's notions of speech acts which, though relevant to the study of communicative context, hardly constitute any dramatic shift of attention in the scope of inquiry.

If we are going to allow a greater role for the communicative context in our study of language, then we need to be open to a larger range of data than we are accustomed to. The language corpora we are in the habit of using do not normally encompass contextual information, e.g., intonation, pauses, eye movement, etc. in the case of conversational exchanges. It is, of course, possible in theory to include such information in a corpus (see, for example, McEnery, Xiao, and Tono 2006: 40-41 for a discussion of models of pragmatic annotation), but such corpora are not widely used. The predominant interest remains in the textual or verbal component of communication, abstracted from context, rather than the communicative activity which is taking place in the context. Hunston (2002: 94) is fairly typical of most linguists (cognitive or otherwise) in terms of her priorities: "However much annotation is added to a text, it is important for the researcher to be able to see the plain text, uncluttered by annotation labels. The basic patterning of the words *alone* must be observable at all times [in the corpus data]." [my italics]

If cognitive linguists are to pursue the study of "situated instances of language use" (cf. Greens and Evans above), then one might reasonably hope for a broadening of the field of inquiry (cf. Wichmann 2007: 82-83 in which the author calls for data from all channels of communication to be included in our corpora). Charles Goodwin in publications such as Goodwin (1979, 1980, 1981) offers a glimpse into what it might mean to study situated instances of language use. He goes much further than we are accustomed to in traditional "conversational analysis", directing attention to the psychological and social processes which occur during conversation: the direction of gaze on the part of participants and when one participant's gaze meets another participant's eyes, withdrawing a gaze, the lowering of the volume of the voice, whether one can be certain that a stretch talk was heard by other participants, etc. This kind of research program represents the study of situated instances of language *par excellence* and cognitive linguists have much to learn from that research.[2]

3. Registers

We can be grateful that, along with increased use of corpora in linguistics, there has been an increased awareness of the relevance of registers, or genres, to the description of linguistic phenomena. One does not need to look further than such well-known corpora as the British National Corpus (BNC), the American National Corpus (ANC), and the International Corpus of English (ICE) to appreciate how widespread the differentiation of genres has become. Within the field of cognitive linguistics, there is, I believe, widespread recognition of the need to separate out genre-specific behaviours, a practice which needs to be maintained if we are concerned with probing more deeply into fine-grained semantic properties of words or constructions. Sometimes, there can be valid reasons for working with an entire corpus such as ICE without distinguishing register. Searching the whole of ICE would be justified, for example, if one is simply exploring all possible contexts of a search term, or all possible nuances carried by a search term, especially at an initial, exploratory stage of an investigation.

A "spoken vs. written" distinction has become so commonplace in corpus work that it seems unremarkable as a major division. However, it is by no means a simple separation and the dangers of too easy an acceptance of such a dichotomy are worth repeating here, even if they are obvious. There is spoken language which derives from fully premeditated, edited, revised, and polished written language designed to be read, e.g., formal speeches delivered to an audience, formal lectures, news broadcasts, etc. The language of such performances has much more in

common with highly formal writing which might appear in the written part of a corpus than it does with more spontaneous forms of oral communication. Conversely, there are very spontaneous forms of written language which share much, stylistically, with conversation, e.g., internet blogs, internet chat room talk, the styles of some personal diaries. Spontaneous conversation would seem to occupy a special place among all the genres in so far as it is represents a relatively basic kind of human interaction.

One does not necessarily have to agree that face-to-face conversation is paramount in terms of our communicative activities (and it may not be for some individuals) to accept that it is an important kind of human activity and deserving of study, especially for cognitive linguists. Arguably, it is spoken, face-to-face conversation which is the genre which needs most to be compared and contrasted with the written genres. It is all the more frustrating, then, that spontaneous, face-to-face conversation remains so underrepresented as a genre in electronic corpora, even in "spoken" corpora. The BNC contains approximately 4.2 million words of such conversation and, as such, remains one of the largest resources for this genre of British English. There are many reasons for this: ethical issues surrounding speakers, difficulty of understanding the exact nature of conversation without access to paralinguistic features, difficulty of transcribing even the textual part of conversation etc. Some spoken corpora may contain little or no truly spontaneous conversation and offer little or no insight into the forces at work in ordinary face-to-face conversation.

Hand in hand with the relative rarity of conversational corpora, many of the tools that have been designed to help linguists work better with electronic corpora do a poor job of searching and retrieving in conversational transcripts. Searching for forms and their contexts in the utterances of a particular speaker is not impossible but remains a complicated business in commonly used corpus tools such as Wordsmith Tools (http://www.lexically.net/wordsmith/, Scott 2004). Retrieving forms and larger contexts including the previous utterance/turn and the following utterance/turn is even more difficult in terms of built-in routines in commonly used software. The CLAN tool used with transcriptions following the CHAT format is probably still the most effective tool for this kind of search and retrieval (and free), even though it has been twenty years since these resources first became available. ELAN (http://www.lat-mpi.eu/tools/elan) also offers convenient separation of tiers for speakers and integration with audio/video, making it very suitable for the study of conversation. The North Carolina Sociolinguistic Archive and Analysis Project, SLAAP (http://ncslaap.lib.ncsu.edu/index.php), provides a unique model for the integration of audio and text in an online environment (for an overview, see Kendall 2007). As well as allowing the user to listen to or download an audio file (in this case, interviews), SLAAP has integrated audio and transcript so that the user can undertake an acoustic analysis of any line of the transcript (corresponding to an unbroken stretch of speech). The acoustic analysis is made possible by Praat scripts which run "on the fly" and no additional software is required on the user's computer. SLAAP offers users a welcome choice of formats for the transcript: vertical format with each successive speaker beginning a new row; column format in which utterances by each speaker are shown in one column; a Henderson Graph format in which the duration of each utterance corresponds to the length of a horizontal line, together with the beginning and end times of the utterance (for more details, see Kendall 2007 and Newman 2009).

There are other ways in which tools may build upon, and be constrained by, a "spoken" vs. "written" dichotomy, potentially leading the researcher, in subtle and unintended ways, away from an interest in face-to-face conversation. BNCWeb (http://www.bncweb.info/) is a case in point. While BNCWeb is an extremely attractive interface for working with the BNC, and one that I would strongly recommend, it nevertheless has certain inbuilt biases which might encourage a researcher to explore (all) spoken vs. (all) written contrasts, rather than, say, face-to-face conversation vs. written genres. The built-in features I have in mind are: (1) BNCWeb includes pre-compiled frequency lists for spoken vs. written, not for face-to-face conversation; (2)

word association measures (MI, z-score etc.) can be obtained reliably for spoken vs. written, not for face-to-face conversation; (3) processing speed is best for all the corpus, spoken sub-corpus, or written sub-corpus. To be fair, one should point out that the BNC itself attaches most importance to the spoken vs. written distinction and, in a sense, BNCWeb is just allowing the user to access the corpus following the inherent major genre division in the BNC. And one must also add that the BNCWeb does allow a user to investigate numerous features of face-to-face conversation as a genre in its own right.

4. Tags

POS tags and syntactic relation markup are basic to some corpus linguistic methods, e.g., collostructional analysis. But neither the parts of speech nor syntactic relations are uncontroversial (neither for English, nor for other languages) and analyses which appeal uncritically to such concepts invite criticism.

The part of speech categories available for English depend upon traditions of analysis and classification, and there are different traditions (different degrees of abstraction, different terms, different degrees of sub-classification, relevance of semantic vs. syntactic criteria in the classification etc.). Arguably, the most important role for linguists should be the critiquing and refinement of these classifications into parts of speech for English and other languages, rather than the acceptance of any one of these systems of parts of speech. Huddleston's (1980) detailed summary of the different structural properties of thirty English auxiliary and modal verbs is still instructive in the manner in which the author demonstrates how each of these verbs behaves in its own unique way with respect to a set of structural properties. Huddleston considers a total of thirty-seven properties which include a number of properties less commonly appealed to in the literature, but which are no less interesting or important because of that, e.g., whether the verb can carry an emphatic stress to make a polarity contrast, whether the verb can occur immediately before a 'deletion' site, and whether the verb can take a negative complement. Huddleston's summary of properties makes it clear that there are many subtle ways in which this set of verbs may be subcategorized. Indeed no two verbs display exactly the same properties. His discussion is a reminder of the important role that linguists have to play in directing attention to the complexities and obstacles associated with part-of-speech classifications. Cognitive linguists should be prepared to *problematize* part of speech systems rather than simply accept them as given. Similar observations hold for the related issue of grammatical relations within a clause and how these are determined. Croft (2001) strikes me as a good example of the kind of original research that cognitive linguists could be carrying out on both parts of speech systems and grammatical relations, in a manner very compatible with the tenets of cognitive linguistics. This is not to say that linguists should never appeal to the traditional categories, only that they should be constantly aware of alternatives and hence cautious about claims based on such categories.

Quite apart from the issue of deciding upon a particular system for parts of speech or grammatical relations, there is the issue of how successfully a tagging algorithm captures the relevant facts. Consider the tags assigned to the *singing* forms in the examples in (2) and (3), taken from the BNC.

(2)	a.	*Cos I kept singing it this morning.*	VVG
	b.	*She says she couldn't stop singing.*	VVG-NN1
(3)	a.	*The singing nearly raised the roof.*	NN1-VVG
	b.	*Sandy can lead the singing.*	NN1

VVG refers to the *–ing* form of a lexical verb, NN1 refers to a singular noun, and a hyphenated tag indicates an ambiguous result with the more likely tag appearing first. In (2a), it is presumably the presence of the direct object *it* immediately after *singing* that contributes to an

unambiguous VVG tag assignment, whereas the absence of any direct object phrase in (2b) leads to the ambiguous tag assignment. It is not clear just what one should make of this difference in tagging. Accept it as a potentially important difference or disambiguate it in favour of one of VVG or NN1? In any case, it is up to the researcher to make the decision; the tagging of the corpus does not in any way "solve" the linguistic problem of determining parts of speech. The tag assignment of NN1-VVG in (3a) is presumably related to the occurrence of the adverb *nearly* immediately after *singing*. *Nearly* modifies the following verb *raised*, not *singing*, in (3a) and should not have any effect on how *singing* is analyzed for its part of speech. A linguist would surely analyze *singing* in an identical way in (3a) and (3b). Again, the tags of a corpus are only a starting point for analysis, not the end point.

The experience of working with markup of grammatical relations reported in Gries, Hampe, and Schönefeld (to appear) highlights the dangers of relying, uncritically, on markup within the corpus. The authors were interested in properties of an "*as*-predicative" in ICE-GB, exemplified in (4):

(4) a. *I do not regard the Delors Report as* [$_{NP}$ *some kind of sacred text*].
b. *Such a force could never be described as* [$_{AdjP}$ *purely deterrent*].
c. *We see the hard ECU as* [$_{S}$ *being extremely useful*].

Grammatical relations can be explored in ICE-GB using the accompanying ICECUP program distributed with the corpus. The authors extracted all occurrences of the relevant search pattern [$_{VP}$ Vcomplex transitive [$_{PP}$ *as*]] using ICECUP. After some manual correction of the output, the authors found a total of 687 tokens of the *as*-predicative construction. As the authors point out, though, this number is misleading, as the search method fails to identify many more instances of this construction in the corpus. By searching for [$_{PP}$ *as*], followed by manual identification of all occurrences of the *as*-predicative, the authors retrieved 1,131 tokens of the construction, a full 65% more than in the first attempt relying upon the "complex transitive" parse in the corpus.

5. Lemma vs. inflected forms

It has been become commonplace, at least in English corpus linguistic research, to report on the behavior of lemmas (e.g. GET), as opposed to the inflected forms that constitute the lemma (e.g., *get* as an infinitive, *get* as a present tense form agreeing with a plural subject, *gets* as a present tense form agreeing with a singular subject, *got* as a past tense form etc.). Recent studies, however, have revealed interesting patterning around particular inflected forms, as opposed to lemmas, suggesting that investigating language at the inflectional level is a promising line of inquiry.[3] In a study of REMEMBER in spoken corpora, Tao (2001, 2003) discusses the prominence of the simple present tense forms of this verb, used with a first person singular subject (*I remember*) or a null subject (*remember*) and suggests a grammaticalization process is under way, confined to particular inflected forms of the verb. Scheibman (2001), in a study of informal conversation, found that 1st singular and 2nd singular subjects occur with particular verbs of cognition with a relative high frequency (*I guess, I don't know, you know, I mean*) reflecting a particular pragmatic value for such combinations in conversation. Scheibman (2001: 84) also emphasizes the need to examine "local" patterns in grammatical research and cautions against relying just on the superordinate grammatical categories (person, verb type, tense etc.). Deignan (2005: 158-159) discusses the different evaluations or "prosodies" that attach to singular *rock* vs. plural *rocks* in the sub-corpora of the Bank of English, reporting a tendency for the singular to have positive evaluations (as in *the rock on which our society is built*) while the plural tends to have negative evaluations (*their marriage has been on the rocks for a while*). Newman and Rice (2004) report on how the inflectional differences between the *–ing* and past tense forms in the

pairs *sitting and.../sat and..., standing and.../stood and…, lying and.../lay and...* influence the range of following verbal collocates. Newman and Rice (2006) explore the different constructional patterns found with positive, comparative, and superlative forms of adjectives.[4] Table 1 summarizes the key results from that study for the pair *slight* and *slightest*. The table lists the top 25 noun collocates in Adjective + Noun combinations for these two forms and the frequency of occurrence of the collocation. The results are sorted according to log-likelihood, as calculated by BNCWeb. The two lists of nouns show different tendencies in their semantic fields. With the positive form *slight*, the majority of nouns in the list have meanings related to 'change, variation', e.g., *increase*, *variation*, *modifications* etc. With the superlative form *slightest*, on the other hand, the majority of nouns have meanings from the domain of 'cognition, perception, intention', e.g., *hint*, *idea*, *doubt* etc. The different semantic tendencies of the noun collocates in these lists can only be appreciated when the researcher is investigating words at the level of the inflected forms. Multivariate analysis techniques which incorporate the full range of related forms of a lemma and a large number of variables associated with these forms are promising techniques for exploring, at least initially, sub-patterns which might well escape detection if we were to rely solely on linguistic intuition to guide us. The clustering analysis underlying "behavioral profiles", as illustrated in Gries (2006), Gries and Divjak (2009), and Gries and Otani (to appear) offer opportunities for exploring differential patterning between inflected forms. Gries and Otani (to appear), for example, reveals intriguing clustering of positive, comparative, and superlative forms of adjectives. Similarly, correspondence analysis, as exemplified in Glynn (2009, to appear), enables a researcher to discern subtle patterning around related word froms involving multiple factors.

INSERT TABLE 1 ABOUT HERE

Some corpus linguistic tools, for example BNCWeb and Sketch Engine (Kilgarriff and Tugwell 2001; Kilgarriff, Rychly, Smrz, and Tugwell 2004), include ways to examine the frequency and contextual information about either inflected forms or lemma. Sketch Engine has a particularly attractive feature of signaling when a particular inflected form has a "salient" distribution. So, for example, the *-ing* (VVG) form *waiting* has a relatively high frequency of occurrence within the set of inflected forms of the verb lemma WAIT. *Waiting* occurs 8,053 times in the BNC, which is 40.62% of all forms of the lemma WAIT which occurs 19,824 times. This puts WAIT in the top 10% of verbs ranked by descending relative frequency of their *–ing* forms (as calculated by Sketch Engine), hence *waiting* is a salient inflected form of WAIT. This key result, that *waiting* occupies a special place among the inflected forms of WAIT, is indicated along with other results from a Word Sketch query without the need for any calculations on the part of the user. Or, to take a noun example, WOMAN, the plural form *women* occurs 34,223 times and accounts for 60.87% of tokens of the lemma WOMAN, putting WOMAN in the top 10% of nouns in terms of relative frequency of their plural forms, hence this fact is indicated in the Word Sketch. The same is not true, for example, of the plural form *men* which accounts for only 37.55% of the forms of MAN. While acknowledging this attractive feature of Sketch Engine, it should be mentioned that the feature is implemented in a fairly limited way, at this point in time. Only the *–ing* form and passive use is reported on in this way for verbs, and only the use of the plural form is reported on for nouns. Furthermore, the "Sketch Differences" tool of Sketch Engine, which allows easy comparison of constructional possibilities of two words based on the BNC, is only implemented with respect to differences between two lemmas–one couldn't compare constructional possibilities of, say, *remember* and *remembered*, which as Tao has shown is a fruitful area of investigation.

6. Quantitative issues

Let us start with the the question of establishing how many words there are in a corpus – something that might appear to some to be too trivial for serious discussion. With so many sophisticated statistical techniques now available for measuring the frequencies and associations within a corpus, one is more used to discussion of much more complex quantitative issues corpus linguistics. However, almost all statistical calculations used in corpus linguistics, from the relatively simple to the most sophisticated, incorporate word counts into their formulae and so establishing the size of a corpus is not inconsequential. Nor is it a trivial task.

Establishing the number of "words" in the BNC is not straightforward. The BNC World Edition is said to contain 100,467,090 orthographic words. One could use this figure as a basis for quantifying the size of the corpus, though the approximation of 100 million words is also used on occasion. But one might also consider using the "w-units" as a basis. "w-units" are the items tagged for parts of speech and may be smaller or larger than what we know as orthographic words in English, as exemplified by the tagging of *out of* and *gonna* in (5).

(5) a. <w type="PRP" lemma="out of">out of </w>
b. <w type="VVG" lemma="gon">gon</w><w type="TO0" lemma="na">na </w>

In (1a), the sequence *out of* is analyzed as a single (compound) preposition, hence one w-unit. The number of compound prepositions and compound conjunctions which are tagged as single w-units is substantial, including forms such as *from time to time* (1,573 tokens), *in addition* (4,229), *in front of* (5,558), *in general* (3,847), *in order* (12,025), *in terms of* (9,238), *more than* (13,149), *out of* (41,845), *of course* (24,091), *rather than* (19,332).[5] Conversely, in (2b), the single word *gonna* is analyzed as two w-units the *–ing* form of a lexical verb GON followed by the infinitival TO. As the analyzed units, assigned a part of speech, the w-units are a perfectly reasonable basis for quantifying the size of the corpus, but this would be different from the number of orthographic words.[6] BNCWeb, in its calculations of statistics, relies on the total number of tagged items in the corpus which includes punctuation marks (Sebastian Hoffman, personal communication). Again, this is not an unreasonable way to proceed–after all, punctuation such as full stops, commas etc. are often seen as providing important contextual clues in understanding the behavior of words in a corpus. Punctuation marks contribute a kind of meaning to a corpus, in this view, and so deserve to be counted. It will be obvious that different decisions about what unit to count affects any further calculations based on total number of units. And, as noted in Footnote 6, even if one decides to count w-units, one is faced with different figures even just consulting the BNC homepage.

A different kind of question arises when carrying out collostructional analysis. In this kind of analysis the researcher arrives at a measure of the statistical significance of the occurrence of a lexeme L in a construction C (cf. Stefanowitsch and Gries 2003: 218). One of the numbers required for this analysis is the frequency of "all other constructions with lexemes other than L" in the corpus. The question arises as to how to count the number of relevant constructions in the corpus, a question considered also by Schmid (this volume). In the case of the construction [N *waiting to happen*], Stefanowitsch and Gries choose to count the total number of verb tags as the number of constructions with lexemes other than L. In the absence of more detailed discussion on this point by the authors, there seem to be a number of alternative ways to proceed when determining the number of constructions to take into account, as pointed out by Schmid. One might prefer to consider *–ing* forms of lexical verbs, for example, given the fact it is specifically the *–ing* form of WAIT which features in this construction. Or one might prefer to consider all instances of verbs concatenated with the infinitive *to*, counting each pair of concatenated verbs as one instance of the construction.

It is inevitable that the quantitative techniques that we use in carrying our corpus-based research will become increasingly sophisticated and a familiarity with quantitative and statistical techniques of analysis will simply be unavoidable. I see myself as an end-user, rather than a creator, of statistical techniques and so for me the challenge that presents itself is one of understanding the strengths and weaknesses of the techniques and choosing between competing techniques. I have not found this particularly easy, in part because even those who are more statistically sophisticated than I am seem to have more questions than answers when it comes to deciding what "best practice" is supposed to be. A case in point is Kilgarriff (2005), published in the first volume of the journal *Corpus Linguistics and Linguistic Theory*, and the commentary on it by one of the editors of the journal, Gries (2005). The two papers address the issue of "null hypothesis testing" in corpus linguistics, and Gries takes up the challenge posed by Kilgarriff's claim that null-hypothesis significance testing (which assumes "randomness") leads to unhelpful or misleading results. Gries (2005: 281) suggests ways to deal with some of the problems raised by Kilgarriff, but concludes with the words: "On the basis of these results, it is very difficult to decide what the right quantitative approach in such word-frequency studies may be. What is not so difficult is to notice that much more exploration in these areas is necessary and that the results show how much we may benefit from taking up methodological proposals from other disciplines to refine, and add to, our methods…". The Gries commentary on the Kilgarriff article typifies for me the conundrum we are faced with when we come to apply more sophisticated statistical measures, most of which assume some kind of randomness in the population (= words in the corpus) being measured. The application of statistical measures in corpus linguistics is clearly still very much in its infancy and, while there are various statistical practices that have become commonplace in corpus linguistics, one cannot assume they represent "best practice" and one must be at least a little circumspect about all such results.

In light of the foregoing remarks, I think we should still concede a place for the "less sophisticated" quantitative measures in our work with corpora. As an example of a relatively unsophisticated approach, one might cite Schmid's measures of *attraction* and *reliance* measures which I find both revealing and useful, even if Schmid himself describes them as "simple arithmetic" (Schmid this volume, section 5.1). These measures, as they apply to nouns are given below:

(6) a. Attraction = $\frac{\text{frequency of a noun in a pattern x 100}}{\text{total frequency of the pattern}}$

b. Reliance = $\frac{\text{frequency of a noun in a pattern x 100}}{\text{total frequency of the noun in the corpus}}$

Attraction measures the extent to which a particular pattern attracts a noun, while reliance measures the extent to which noun appears in one particular pattern versus other patterns. Schmid (this volume) provides a helpful discussion of the strengths and weaknesses of these formulae and an insightful comparison of these measures with the results of collostructional analysis. Given the controversies surrounding the use of statistical measures in corpus linguistics, it seems perfectly reasonable to continue to make use of a range of more qualitative techniques to understand the behavior of words in context. Hunston (2002) helpfully describes a variety of useful methods for examining usage of words in a corpus, working from concordance lines. While some of the methods that she describes rely upon statistical measures such as MI, z-score, and t-score, others do not. The non-statistical approaches that she discusses are still properly called "methods" which have their own rationale and rigour. And where the researcher is exploring subtle semantic nuances of a word as evidenced in the context, including the larger discourse context, then one has little choice but to work with these "softer" quantitative methods. The chapters in Deignan (2005) and Stefanowitsch and Gries (2006), for example, contain many insightful and thought-

provoking analyses of corpus data employing little more than percentages and chi-square measures.

7. Conclusion

If we accept that usage is to play a key role in cognitive linguistic research (as claimed in Evans and Green 2006), then corpora and the analytical techniques associated with them will necessarily be an ever more important focus for researchers in this area. I find myself in complete agreement with Geeraerts (2006:45), then, when he writes that "a program in Cognitive Linguistics worthy of that name should resolutely opt for advanced training in empirical linguistics", where advanced training in empirical linguistics necessarily includes advanced training in statistical techniques.

There is an abundance of corpora available to researchers, and so one can look forward to much more corpus-based research into cognitive linguistics. Corpus data seem very accessible and "ready to use", especially with some of the corpus tools now available to researchers. But it is important to appreciate alternative ways of proceeding with corpus-based research. In the discussion above I have emphasized the need to bear these alternatives in mind in working with a corpus. I believe that the alternatives I have reviewed in the discussion above are all worth pursuing further: working with language as structure *and* language as communicative activity; the inclusion of conversational language *and* other genres in studies of language; investigating both inflected forms *and* lemmas; a tolerance for both quantitative *and* qualitative methods. Our goal as cognitive linguists should be one of maintaining a balance between alternative kinds of evidence and alternative research methodologies.

Notes

[1] The increasing interest in usage-based data is true of the field of linguistics and is by no means restricted to cognitive linguistics. An Editor of *Language* has observed, with reference to the contents of the journal: "…we seem to be witnessing…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." (Joseph 2004: 382)

[2] One should mention, too, the contributions of Harris (1996, 1998) who argues forcefully for an agenda for the study of language which situates language well and truly in the context of communication, what Harris calls an "integrationist approach". Thorne and Lantolf (2006) is a more recent statement of a similar kind in which the authors argue for a "Linguistics of Communicative Activity", the goal of which is to "disinvent language understood as an object and to reinvent language as *activity*..." (Thorne and Lantolf 2006: 171, italics original).

[3] Stubbs (2001: 99) draws attention, in passing, to the issue of investigating different inflected forms, as opposed to lemmas, and cites a remark by Sinclair (1991: 8) suggesting the value of investigating the different forms of a lemma in terms of their uses. Further discussion of the role of inflected forms vs. lemmas can be found in Newman (2008).

[4] Claridge (2007), investigating the superlative of English adjectives in conversational English, demonstrates the value of considering specific inflected forms of adjectives, as opposed to lemmas. She reports on a range of properties of the superlative which could not be claimed of the class of adjectives in general.

[5] The tagging system can lead to difficulties when it comes to retrieving some of these forms. Mark Davies comments on this in connection with searches for multiword units like *in charge of* when using his Contemporary Corpus of American English (COCA) site: "…the multi-word units work only with queries that involve at least one specified word in the query. In other words, [* charge *] will find *in charge of*, but a query composed of just wildcards [* * *] or parts of speech ([prp] [nn1] [pr*]) would not" (http://corpus.byu.edu/bnc/). Davies' comment is a

reminder of just how careful one must be in retrieving information from the BNC. The user needs to be well aware of the full range of multiword w-units in performing searches.

[6] On the relevant web page of the BNC, the number of w-units in the whole corpus is said to be "slightly less than" 97,619,934. On the same page, one paragraph after this number is given, we are told that the total number of w-units is 89.30 (written) + 10.58 (spoken) = 99.88 million, a discrepancy of more than 2 million words (http://www.natcorp.ox.ac.uk/corpus/index.xml.ID=numbers, accessed 3 July 2007) The BNCWeb, in its reporting on searches in the whole corpus, gives the number 97,626,093, which presumably refers to w-units, rather than orthographic words, but still differs from the two numbers given on the BNC homepage.

References

Chomsky, Noam 1965 *Aspects of the Theory of Syntax.* Cambridge, Mass.: M.I.T. Press.

Claridge, Claudia 2007 The superlative in spoken English. In Roberta Facchinetti (ed.), *Corpus Linguistics 25 Years On*, 121-148. Amsterdam/New York: Rodopi.

Croft, William 2001 *Radical Construction Grammar.* Oxford: Oxford University Press.

Deignan, Alice 2005 *Metaphor and Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins.

Dirven, René and MarjolijnVerspoor 2004 *Cognitive Exploration of Language and Linguistics*. 2nd ed. Amsterdam: John Benjamins.

Evans, Vyvyan and Melanie Green 2006 *Cognitive Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.

Geeraerts, Dirk 2006 Methodology in cognitive linguistics. In Gitte Kristiansen, Michel Achard, René Dirven, and Francisco J. Ruiz de Mendoza Ibáñez (eds.), *Cognitive Linguistics: Current Applications and Future Perspectives*, 21-49. Berlin: Mouton de Gruyter.

Glynn, Dylan 2009 Polysemy, syntax, and variation: A usage-based method for cognitive semantics. In Vyvyan Evans and Stéphanie Pourcel (eds), *New Directions in Cognitive Linguistics,* 77-104. Amsterdam/Philadelphia: Benjamins.

Glynn, Dylan to appear Synonymy, frames, and fields: Developing usage-based methodology for cognitive semantics. In Hans-Jörg Schmid and Susanne Handl (eds.), *Cognitive Foundations of Linguistic Usage Patterns: Empirical Studies*. Berlin: Mouton de Gruyter.

Goodwin, Charles 1979 The interactive construction of a sentence in natural conversation. In G. Psathas (ed.), *Everyday Language: Studies in Ethnomethodology*, 97-121. New York: Irvington.

Goodwin, Charles 1980 Restarts, pauses, and the achievement of mutual gaze at turn-beginning. *Sociological Inquiry* 50 (3-4): 272-302. (Special Double Issue on Language and Social Interaction, edited by Don Zimmerman and Candace West).

Goodwin, Charles 1981 *Conversational Organization: Interaction between Speakers and Hearers.* New York: Academic Press.

Gries, Stefan Th. 2005 Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory* 1 (2): 277-294.

Gries, Stefan Th. 2006 Corpus-based methods and cognitive semantics: The many meanings of *to run*. In Stefan Th. Gries and Anatol Stefanowitsch (eds.*), Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*, 57-99. Berlin/New York: Mouton de Gruyter.

Gries, Stefan Th. and Dagmar S. Divjak 2009 Behavioral profiles: A corpus-based approach towards cognitive semantic analysis. In Vyvyan Evans and Stéphanie Pourcel (eds.), *New Directions in Cognitive Linguistics*, 57-75. Amsterdam/Philadelphia: John Benjamins.

Gries, Stefan Th., Beate Hampe, and Doris Schönefeld to appear Converging evidence II: More on the association of verbs and constructions. In Sally Rice and John Newman (eds.), *Experimental and Empirical Methods in the Study of Conceptual Structure, Discourse, and Language*. Stanford, CA: CSLI.

Gries, Stefan Th. and N. Otani to appear Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal*.

Harris, Roy 1996 *Language and Communication: Integrational and Segregational Approaches.* London: Routledge.

Harris, Roy 1998 *Introduction to Integrational Linguistics*. Oxford: Elsevier Science Ltd.

Huddleston, Rodney D. 1980 Criteria for auxiliaries and modals. In Sidney Greenbaum, Geoffrey N. Leech and Jan Svartvik (eds.), *Studies in English Linguistics for Randolph Quirk*, 65-78. London/New York: Longman.

Hunston, Susan 2002 *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Joseph, Brian D. 2004 The Editor's Department: On change in *Language* and change in *language*. *Language* 80 (3): 381-383.

Kendall, Tyler 2007 Enhancing sociolinguistic data collections: The North Carolina Sociolinguistic Archive and Analysis Project. *University of Pennsylvania Working Papers in Linguistics* 13 (2): 15-26.

Kilgarriff, Adam 2005 Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1 (2): 263-276.

Kilgariff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell 2004 The Sketch Engine. *Proceedings of EURALEX 2004*, Lorient, France, 105-116.

Kilgarriff, Adam and David Tugwell 2001 WORD SKETCH: Extraction and display of significant collocations for lexicography. *Proceedings of the Collocations Workshop, ACL 2001*,Toulouse, France, 32-38.

McEnery, Tony, Richard Xiao, and Yukio Tono 2006 *Corpus-based Language Studies: An Advanced Resource Book.* Milton Park: Routledge.

Newman, John 2008 Aiming low in linguistics: Low-level generalizations in corpus-based research. *Proceedings of the 11th International Symposium on Chinese Languages and Linguistics (IsCLL-11)*, May 23-25 2008, National Chiao Tung University, Hsinchu, Taiwan. Available at http://johnnewm.jimdo.com/downloads.php.

Newman, John and Sally Rice 2004 Patterns of usage for English SIT, STAND, and LIE: A cognitively-inspired exploration in corpus linguistics. *Cognitive Linguistics* 15: 351-396.

Newman, John and Sally Rice 2006 English adjectival Inflection: A radical Radical Construction Grammar Approach. Conceptual Structure, Discourse, and Language Conference, University of California, San Diego, November 5 2006.

Scheibman, Joanne 2001 Local patterns of subjectivity in person and verb type in American English conversation. In Joan L. Bybee and Paul Hopper (eds.), *Frequency and the Emergence of Linguistic Structure*, 61-89. Amsterdam/Philadelphia: John Benjamins.

Schmid, Hans-Jörg this volume Does frequency in text really instantiate entrenchment in the cognitive system? And do we have a quantitative grip on either of them?

Scott, Mike 2004 *WordSmith Tools* Version 4, Oxford: Oxford University Press.

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

Stefanowitsch, Anatol and Stefan Th. Gries (eds.) 2006 *Corpus-Based Approaches to Metaphor and Metonymy*. Berlin/New York2: Mouton de Gruyter.

Stubbs, Michael 2001 *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Tao, Hongyin 2001 Discovering the usual with corpora: The case of remember. In Rita C. Simpson and John M. Swales (eds.), *Corpus Linguistics in North America: Selections from the 1999 Symposium*, 116-144. Ann Arbor: University of Michigan Press.

Tao, Hongyin 2003 A usage-based approach to argument structure. *International Journal of Corpus Linguistics* 8: 75-95.

Teubert, Wolfgang 2005 My version of corpus linguistics. *International Journal of Corpus Linguistics* 10 (1): 1-13.

Thorne, Steven L. and James P. Lantolf 2006 A linguistics of communicative activity. In Sinfree Makoni and Alastair Pennycook (eds.), *Disinventing and Reconstituting Languages*, 170-195. Clevedon: Multilingual Matters.

Tummers, José , Kris Heylen, and Dirk Geeraerts 2005 Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory* 1 (2): 225-261.

Wichmann, Anne 2007 Corpora and spoken discourse. In Roberta Facchinetti (ed.), *Corpus Linguistics 25 Years On*, 73-86. Amsterdam/New York: Rodopi.

slight N	Frequency	LL	*slightest* N	Frequency	LL
increase	57	499	***hint***	25	338
smile	48	455	*bit*	36	303
variations	28	284	*movement*	29	271
angle	27	274	***idea***	25	203
improvement	28	256	***doubt***	22	198
modifications	20	234	***sign***	20	192
acidity	17	231	*chance*	21	184
breeze	20	215	***interest***	24	181
change	34	213	***intention***	15	152
delay	20	190	*provocation*	10	145
frown	15	186	***suspicion***	12	142
difference	26	183	*difference*	14	115
shrug	14	181	*thing*	17	109
changes	29	181	***inclination***	8	106
bow	16	168	*touch*	11	93
decrease	15	167	*breeze*	8	91
exaggeration	13	165	*flicker*	6	79
figure	26	162	***inkling***	5	75
pause	16	160	***trace***	7	71
differences	20	145	***notice***	9	70
fall	18	142	*degree*	9	68
movement	22	140	*sound*	9	62
rise	19	136	***nod***	5	58
variation	15	131	***suggestion***	6	55
flush	10	120	***encouragement***	5	51

Table 1. Top 25 noun collocates of *slight* and *slightest* in the whole BNC, with frequencies of collocation, sorted by log-likelihood (LL) value, as calculated by BNCWeb. Words in bold show the preference for 'change' nouns with *slight* and 'cognition, awareness' nouns with *slightest*.