Corpus linguistics, theoretical linguistics, and cognitive/ psycholinguistics: Towards more and more fruitful exchanges

Stefan Th. Gries

University of California, Santa Barbara

Abstract

This article discusses my version of corpus linguistics, its relation to what I think are neighboring fields (mainly cognitive and psycholinguistics), how corpus linguistics can and should enter into more mutually beneficial relations with these fields, and the from my perspective most promising contemporary cognitive and psycholinguistic approach that provides a natural point of connection to corpus-based approaches.

"Resistance is futile!" (The Borg)

1. Corpus linguistics and (more) theoretical approaches

The relation between corpus linguistics (CL) and linguistic theory has traditionally been somewhat problematic. I think there are several reasons for this: corpus linguists differ as to what they think CL is: a tool, method(ology), discipline, theory, paradigm, framework; there are some things that make CL appear less attractive to the observer from theoretical linguistics; and some corpus linguists have a rather inflexible gate-keeping attitude towards what the field is (supposed to look) like that, ultimately, impedes progress rather than advances it.

1.1 Within corpus linguistics: What we think corpus linguistics is and has been

As for the former, some consider CL a theory, for instance Leech (1992: 106), Stubbs (1993: 2f.), Tognini-Bonelli (2001: 1), Teubert (2005: 2). Others consider CL a methodology, such as McEnery and Wilson (1996), Meyer (2002), Bowker and Pearson (2002), McEnery, Xiao and Tono (2006: 7f.), Hardie (bcd).[1] The latter two positions are particularly worth quoting here:

> corpus linguistics is a whole system of methods and principles of how to apply corpora in language studies and teaching/learning, it certainly has a theoretical status. Yet theoretical status is not theory in itself. (McEnery, Xiao and Tono 2006: 7f.)

> As a corpus linguist I consider myself primarily a methodologist and CL primarily a methodology, to be applied to whatever theory seems most appropriate for the task at hand. (Hardie, bcd)

Yet other corpus linguists (e.g. Aarts 2002; Teubert 2005; Williams 2006) use even other labels such as *discipline* or what I would call a methodological commitment: "[CL] is rather an insistence on working only with real language data taken from the discourse in a principled way and compiled into a corpus" (Teubert 2005: 4).

It is probably worth pointing out that whether scholars attribute the status of theory to CL or not often somewhat coincides with where they are on the continuum of corpus-driven and corpus-based linguistics. Corpus-driven linguists

- aim to build theory from scratch, completely free from pre-corpus theoretical premises;
- base theories exclusively on corpus data;
- often reject corpus annotation (as a pre-corpus theoretical commitment), going so far so as to resort to absurdly unfounded generalizations such as "[c]ognitive linguists like to work with annotated language data. But annotations presume presupposed categories not validated by corpus evidence" (Teubert 2010: 355) – Teubert *knows* that all – all! – annotation cognitive linguists use is not validated by corpus evidence??

Corpus-driven linguistics, in essence, means 'bottom-up'. The following quote by Teubert (2005: 4) is instructive in this context:

> While corpus linguistics may make use of the categories of traditional linguistics, it does not take them for granted. It is the discourse itself, and not a language-external taxonomy of linguistic entities, which will have to provide the categories and classifications that are needed to answer a given research question.

Corpus-based linguistics, on the other hand, approaches corpus data from the perspective of moderate corpus-external premises, with the aim of testing and improving such theories, and often uses corpus annotation.

I'm more of a corpus-based linguist and, agreeing with Hardie and McEnery's (2010) stance, consider CL "a major methodological paradigm in applied and theoretical linguistics" Gries (2006a: 191). Why? Here I will mention two reasons… First, I agree with Jan Aarts, who, although he coined the term *corpus linguistics*, also is

> reported as commenting that the term was coined with some hesitation 'because we thought (and I still think) that it was not a very good name: it is an odd discipline that is called by the name of its major research tool and data source'. (Taylor 2008: 179)

Put differently, I don't accord CL the status of a theory just as I don't think there is a linguistic theory called experimental linguistics or self-paced reading time linguistics even though, just like results from CL, results from self-paced reading times may call into question units/structures/processes assumed in the kind of formal linguistics that (some of) CL was a reaction against.

Second, with the exception of, maybe, Sinclair and Mauranen's (2006) Linear Unit Grammar, I have yet to see what I consider a truly corpus-driven approach, at least within corpus linguistics (cf. below). Even corpus-linguistic studies that consider themselves corpus-driven are often not as corpus-driven as they could or claim to be. This can be seen on at least three different levels. First, from a theoretical perspective on the lexical/grammatical level, a truly corpus-driven approach would, strictly speaking, require a complete distributional analysis of the corpus (with maybe some machine-learning algorithm, neural network, or recursive association rules approach) to initially identify the linguistic units manifested in the data (similar, but more advanced to, say, C.C. Fries's well-known approach). And while some corpus linguists make statements to that effect (cf. Teubert's "Corpus linguists still don't know what a morpheme, a word, a phrase or a pattern is", bcd),

- many corpus-driven studies at least start out from the notion of a word;
- Bill Louw (bcd) has studied *all sorts of*, but my guess is he has not used a bottom-up or even replicable algorithm such as Kita et al.'s (1994) cost criteria, lexical gravity, or some other *n*-gram statistic to identify *all sorts of* as a unit – he has decided that *all sorts of* is a unit on the basis of corpus-external or pre-corpus criteria;
- POS are not uncommon in so-called corpus-driven studies (cf. Xiao 2009: 995; Linear Unit Grammar at least starts out without POS);
- even Halliday (2005: 174), revered by many corpus-driven linguists, writes "a corpus-driven grammar is not one that is theory-free," referring to Hunston and Francis (2000) and Tognini-Bonelli (2001).

Xiao (2009: 995) summarizes the problems of corpus-driven linguistics persuasively: "applying intuitions when classifying concordances may simply be an implicit annotation process, which unconsciously makes use of preconceived theory", and this implicit annotation is "to all intents and purposes unrecoverable and thus more unreliable than explicit annotation".

The second level on which corpus driven work is often less corpus-driven as is assumed is concerned with the more concrete perspective of the lexical/grammatical level. For example, many corpus-driven studies look at *n*-grams, where *n* is arbitrarily defined as one number (currently, *n*=4 is en vogue), but most studies do not

- check whether that number is indeed the best number for all *n*-grams; as one of rather few laudable exceptions, Biber (2009) checks for 5-grams;
- check whether it would not indeed be better to have different *n*'s for different *n*-grams: obviously, 3-grams may miss *according to*, 4-grams may miss *in spite of*, 5-grams may miss *on the other hand*, 6-grams may

miss *as a matter of fact*, and 7-grams may miss *be that as it may* (cf. Kita et al. 1994; Mason 2006; as well as Gries and Mukherjee 2010 on varying-length *n*-grams);
- even consider discontinuous *n*-grams, e.g., those constituted by idioms such as *run a*/*the risk* (e.g. Nagao and Mori 1994).

The third and final level is concerned with the notion/level of the register. Of course, corpus linguists of all persuasions often study words and constructions in corpora that come with divisions into modes (speaking vs. writing), registers, or sub-registers. However, for any given phenomenon, the register distinctions that a researcher chooses to take into account – often out of convenience – may not be most useful from a truly bottom-up perspective, which is why more seriously corpus-driven/bottom-up approaches are required (cf. Gries 2006b, 2010b; Gries et al. 2011).

With regard to this division of corpus-driven vs. corpus-based linguistics, Xiao (2009) states "the distinction between the two is overstated" and that "the corpus-based approach is better suited to contributing to linguistic theory". As to the former, I disagree – if anything, it is *under*stated given that truly corpus-driven work is probably a myth at best. However, to some degree I can understand how, at the time, the issue of corpus-based vs. driven was a useful distinction to raise awareness of the larger issues involved in, and following from, that distinction (Tognini-Bonelli, personal communication). The latter, I find interesting because in effect it says that corpus-driven linguistics, where scholars use corpus-driven characteristics to argue for corpus linguistics as a theory, is in fact less suited to contributing to linguistic theory than corpus-based linguistics, which often views corpus linguistics as a method(ology) 'only'.

1.2 From within corpus linguistics: What we think/say we and others do

Turning to what we and scholars from other disciplines see from, and think about each other, I think there are some things that make corpus linguistics less attractive to the observer from theoretical linguistics. These include some corpus linguists' rather unusual
- ideas about potentially relevant neighboring disciplines;
- ways of defending their perspective(s);
- sometimes narrow and prescriptive views about the nature of the discipline (above and beyond the above issues).

I don't have the space to discuss all of these issues in detail so some examples must suffice. As for (i), Teubert (2010: 395) characterizes the relation between NLP and cognitive linguistics (CogLing) as follows: NLP is one of CogLing's "illegitimate offspring[s]". This statement is, I regret to say, simply completely absurd. I fail to see any connection between these fields other than that Teubert does not want CL to be like either of them, and neither do, I think, most members of these two disciplines. I don’t see cognitive linguistics papers in *Computational Linguistics* or the many different proceedings in which computational linguists publish, nor do I see NLP papers in *Cognitive Linguistics*. As another example, Teubert (2010: 398) flatly asserts, brushing aside 30 to 40 years of psycholinguistic research on speech production, “[w]hen we speak, we do not turn thought into language”. As a final example, consider Mason’s (2007: 2) argument in favor of Linear Unit Grammar (in an otherwise very interesting paper):

> Formal approaches to the description of sentence structure furthermore take for granted a hierarchical (phrase) structure […]. However, language is not produced in that way, but instead is a linear sequence created in stops and starts. A hierarchical structure thus cannot account for the fact that the beginning of an utterance is already produced before the whole sentence has been completely worked out. Similar issues apply for the reception of language. Unlike the hierarchical, a linear approach is more closely related to the way most language is received. Processing usually begins before a complete sentence has been heard or read, and quite often the remaining parts of a sentence can be predicted with high accuracy before its completion.

Not only does this excerpt appear to assume that we do turn thoughts into language, *pace* Teubert, it also betrays a serious misunderstanding of psycholinguistic approaches to language production and comprehension: an incremental approach to language production and comprehension of the type that Mason’s last sentence appears to represent does by no means require abandoning a largely hierarchical view of language structure (cf. Hawkins’s (1994, 2005) or Kempen’s (passim) work on incremental production). A looser definition of constituency may be useful to increase the range of units that are manipulated in comprehension and production to include (linear) multi-word units etc., but that does not mean such units cannot still be analyzed hierarchically.

As for (ii), discourse with and about (more) theoretical linguistics is often characterized by a strange us vs. them gate-keeping warfare that

- involves people’s inability to read, as when multiple commentators in the bootcamp discourse (examples include Wynne and of course Louw) fail to understand from my postings on the corpora list or the publicly posted syllabus of the bootcamp that maximally one sixth of the bootcamp was devoted to statistical tests;
- argues against strawmen, as when Teubert states that “[l]inguistics is the study of real, human language, not the development of useful gadgets simulating the use of language”, something that I think just about every participant in the bootcamp discourse (including me) would subscribe to; in fact, in a recent interview (Gries, forthcoming), I argued that precisely

for such reasons the term *computational linguistics* is, to my mind, sometimes used inappropriately:

- "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 (a useful gadget, I presume) falls under the heading of natural language processing, but I would not call it ___ *linguistics* (even if frequency data from corpora are used to tweak how the machine parses its input), but, if pressed, would call it *linguistic computing*".
- Similarly, when Williams (2010: 403) states, "[a]s Lou Burnard and Chris Tribble so ably pointed out, those who teach and translate are not to be looked down on", I am wondering who this is an argument against – nobody defended a position where teachers and translators are to be looked down on. Finally, Williams' (2010: 402) agreement with "Firth who refuted all mentalism in favour of building models from what we really see rather than from introspective 'knowledge' of what we think might be happening", 'argues' against only a ridiculous caricature of what cognitive linguistics of the kind I have been and will be arguing for is like (cf. below Sections 2.1 and 3).
- uses geographical labels in place of arguments (as when agendas are simply labeled as "transatlantic", which I guess means 'bad');
- contrasts (good) old-fashioned Sinclairian core corpus linguistics with those who "piss into" Sinclair's canonical corpus linguistics tent,[2] who use corpora in "a seemingly inappropriate, toolbox-like, inherently non-Sinclairian way" (Mukherjee, bcd, characterizing the viewpoint he opposes);
- couches interdisciplinary discourse in terms of "hijacking" (Teubert 2010: 356) and "takeover bid[s]" (Williams 2010: 407) or Wynne's condescending (2010: 427) "people who *think* that they are doing corpus linguistics" (my emphasis).

Not only has this kind of discourse never helped anything, but it can also, as Hardie and McEnery argue persuasively, effectively constitute re-writing the history books in how "it elides the methodologist tradition completely from history, and describes the field of corpus linguistics as if it rested solely on the accomplishments of neo-Firthian scholars and the philosophers from whom they draw inspiration" (Hardie and McEnery 2010: 388); Mukherjee's discussion of this type of "corpus dogmatism" is similarly right on target.[3]

As for (iii), some gatekeepers of corpus linguistics have a narrower view of the field, or a more prescriptive attitude towards the field, than the actual and healthily diverse field would appear to support. These are some examples from Teubert:

- “corpus linguistics looks at phenomena which cannot be explained by recourse to general rules and assumptions” (Teubert 2005: 5) – well, I know many corpus linguists who are interested in explaining phenomena this way, esp. since “general rules and assumptions” do not rule out probabilistic rules and assumptions.
- “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” (Teubert 2005: 10) – this comment and others reveal what Mukherjee (2010: 373) characterizes as a “fixation” on meaning in discourse at the cost of meaning in the mind (although everything in discourse was at one point of time filtered through at least one mind; cf. below). And seeing if there is more evidence in the discourse about a lemon’s sweetness appears to me as something for the hypothetical *Journal of Taste and Smell Research*, not the hypothetical *Journal of Corpus Linguistics*.
- Teubert (2005: 2f.) states that “[c]orpus linguistics looks at language from a social perspective. It is not concerned with the psychological aspects of language” (my emphasis), but on the other hand, he writes (ibid.: 7): “Linguistics is not a science like the natural sciences whose remit is the search for ‘truth’. It belongs to the humanities, and as such it is a part of the endeavour to make sense of the human condition. Interpretation, and not verification, is the proper response to the quest for meaning”.

Not only do I not see how blanking out the very thing that makes us human – mind/*Geist* – helps in the endeavour to make sense of the human condition, I will also outline below many ways in which cognitive approaches to language are not only compatible with much recent work in corpus linguistics, but also provide a framework into which corpus-linguistic results can be integrated elegantly.

1.3 Taking stock…

Now, all of this must not distract from the facts that CL in its present form is a young discipline, but has left quite a mark on linguistics in general and theoretical linguistics in particular. However, I think CL can benefit from more interaction because many take the above delimitation(s) of the field too literally and often develop tools/methods that may appear useful when applied with the we-never-talk-about-anything-other-than-the-discourse(s) perspective but that hardly get validated against anything outside the discourses.

For example, there are 20+ measures of dispersion but few corpus linguists try to determine which are best in which circumstances (exceptions include Lyne 1985 and Gries 2008, 2009). For example, there are many different ways to generate *n*-grams, but few corpus linguists try to determine which of these ways result in something corresponding to something outside of the narrow confines of the discourses. For example, there are 30-something measures of collocational strength, but not only do few corpus linguists try to determine which are best when (Evert and Krenn 2005 and Wiechmann 2008 are laudable exceptions), there are now also corpus linguists who pretty much argue for trying different ways to modify existing measures and pick whatever yields results that intuitively (!) appear best and then sell that functionality as part of an unvalidated commercial web-based package. These facts are troubling because such validations are so necessary as studies differ with regard to which, say, measures of attraction yield the best results: Krug (1998) finds string frequency to be most predictive; Gries et al. (2005, 2010) find $p_{\text{Fisher-Yates}}$ to yield good results; Wiechmann (2008) gets the best results with (the theoretically maybe problematic measure) *Minimum Sensitivity* and $p_{\text{Fisher-Yates}}$. Thus, do we as corpus linguists just go on using *MI* (or *t* or …) just because we're supposed to focus on the discourse only and because the WordSketch engine makes that easy? Don't we care there are psycholinguistic results available that bear on our choice of statistics?

Obviously, I think we should and, thus, CL would benefit from applying corpus methods outside of CL and its discourses proper because that would increase CL's visibility in the field of linguistics as a whole and in particular with disciplines that have often independently arrived at similar findings or conclusions, but also because external validation would streamline corpus-linguistic research enterprises. In fact, Butler (2004) argues for a "greater awareness in corpus linguistics of the need for a more powerful and cognitively valid theory" (Hoey 2005: 7). However, if that is so, which theory should CL turn to?

By now it has become obvious that I disagree with most of Teubert's opinions, which is why one can turn to him to guess which theory I have in mind for CL. Here are some instructive quotes:

- "For me, corpus linguistics and cognitive linguistics are two complementary, but ultimately irreconcilable paradigms" (2005: 8).
- "Corpus linguistics localises the study of language, once again, firmly and deliberately, in the *Geisteswissenschaften*, the humanities" (2005: 13).
- "Corpus linguistics looks at language from a social perspective. It is not concerned with the psychological aspects of language" (2005: 2f.).

Adding up all this brings me to a psycholinguistically informed, (cognitively-inspired) usage-based linguistics which should be located, firmly and deliberately, in the social/behavioral sciences.[4] And in some sense, that is the logical choice. First, as we are talking about the humanistic perspective and the *Geisteswissenschaften*, is not illuminating the cognitive system(s) that ultimately give rise to discourse(s) telling us much more about the 'human condition' than interrelations between text files? We can't seriously be in the *Geisteswissenschaften* if the one thing we *a priori* blank out is *Geist*?!

Second, at some point in time, going cognitive is necessary: things only enter into discourse when a speaker has processed them and 'decided' to utter them and, thus, make them part of the discourse, and the way a hearer processes input is also determined by that hearer's internal structure. As Maxwell (bcd) put it:

> I would have thought that meaning was not inherent in any corpus, nor in some community's use of language, but could only be understood (bad term, but I can't think of another) with reference to the individual minds of the people using that language (cf. Washtell, bcd, for a similar statement)

Thus, a psycho- and cognitive-linguistically informed usage-based linguistics it is. But how does this field relate to what's happening in CL?

2. Corpus linguistics and one particular (more) theoretical approach

2.1 Corpus linguistics and cognitive linguistics/psycholinguistics: Some commonalities

If one takes a look at some such frameworks (cf. Gonzálvez-García and Butler (2006) for an excellent discussion of different cognitive/functional models), many commonalities between CL and (newer) developments in psycholinguistically informed, (cognitively-inspired) usage-based linguistics emerge. In fact, many notions and results in CL not only have immediate psycholinguistic and/or cognitive-linguistic relevance, but can also be explained in a more illuminating way once we open our eyes to and explore the large body of evidence that other disciplines have to offer.

For example, when we corpus linguists talk about token frequencies,

- cognitive linguists become interested because, on the whole, token frequencies correlate with degree of entrenchment (Schmid 2000) or phonetic reduction and development of new forms (Fidelholtz 1975);
- psycholinguists become interested because, on the whole, token frequencies correlate with ease/earliness of acquisition (Casenhiser and Goldberg 2005); lexical decision tasks, word naming, picture naming (Howes and Solomon 1951; Forster and Chambers 1973).

When we in corpus linguistics talk about type frequencies,

- cognitive linguists become interested because type frequencies are correlated with (morphological) productivity and language change (type frequency: Bybee 1985; rule reliability: Albright and Hayes 2003);
- psycholinguists become interested because type frequencies are correlated with the productivity of, say, constructions in first/second language acquisition.

When corpus linguists talk about dispersion, which they do too rarely, cognitive and psycholinguists become interested because dispersion has implications for psycholinguistic experiments (Gries 2009) and learning/ acquisition (cf. Simpson and Ellis 2005; Ambridge et al. 2006; Schmidtke-Bode 2009).

When corpus linguists argue against a strict separation of syntax and lexis, cognitive linguists agree, and many psycholinguists have long assumed that words and syntactic patterns are represented as qualitatively similar nodes in a network where, in production, lexical and syntactic nodes are activated when they fit the semantic/pragmatic meaning to be communicated.

When corpus linguists talk about the Idiom Principle ("a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (Sinclair 1991: 110), cognitive linguists become interested because it reminds them of Langacker's (1987: 57) unit, which is

> a structure that a speaker has mastered quite thoroughly, to the extent that he can employ it in largely automatic fashion, without having to focus his attention specifically on its individual parts for their arrangement […] he has no need to reflect on how to put it together

as well as Langacker's (1987: 42) rule-list fallacy, which states

> [t]here is a viable alternative: to include in the grammar both the rules and instantiating expressions. This option allows any valid generalizations to be captured (by means of rules), and while the descriptions it affords may not be maximally economical, they have to be preferred on grounds of psychological accuracy to the extent that specific expressions do in fact become established as well-rehearsed units. Such units are cognitive entities in their own right whose existence is not reducible to that of the general patterns they instantiate.

When corpus linguists talk about words and patterns, psycholinguists become interested because when something attains unit status it can prime and be primed (both lexically and syntactically), and cognitive linguists become interested because Hunston and Francis's (2000) patterns are very similar to Goldberg's (2006) constructions. Compare the following two widely-cited quotes:

> The patterns of a word can be defined as all the words and structures which are regularly associated with the word and contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it. (Hunston and Francis 2000: 37)

> Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully

> predictable as long as they occur with sufficient frequency. (Goldberg 2006: 5)

When corpus linguists talk about concordances, collocations, *n*-grams, colligations – i.e., anything having to do with co-occurrence information – psycholinguists become interested because such co-occurrence information

- helps children discern phonotactic patterns (Saffran et al. 1996);
- can predict reading times (MacDonald 1993) and gaze duration (McDonald et al. 2001);
- helps subjects recognize frequent 4-grams faster (when 1-gram and 2-gram frequency is controlled) (Snider and Arnon, forthcoming);
- language production and comprehension have been shown to be highly item-specific, which is just another way of saying context-bound (e.g. lexically-specific reduction or priming effects).

In a twist whose irony is probably underappreciated, many of these approaches are more corpus-driven than much self-proclaimed corpus-driven work, as when Reddington et al. (1998) and Mintz et al. (2002) apply bottom-up cluster algorithms to corpus data to explain children's recognition of parts of speech.

Also, in conformity with Teubert's social perspective, Croft (2009) and others have been arguing for a cognitive sociolinguistics and the first papers, volumes, and conferences focusing on such issues can now be found. The theory of Construction Grammar even explicitly allows for such a connection: "the function pole in the definition of a construction indeed allows for the incorporation of factors pertaining to social situation, such as e.g. register" (Goldberg 2003: 221). There is also increasingly more work in CogLing relying on corpora. There were several theme sessions on corpora and/or frequency effects in cognitive linguistics at cognitive linguistics conferences, and more than half of all papers in the next proceedings of the American version of the CogLing conference use corpora. Similarly, more and more psycholinguistic work utilizes corpora, as can be seen by searching for the words *corpus*/*corpora* on, say, the website of the *Journal of Memory and Language*.

2.2 Taking stock again…

Given the above, it becomes clear that CL is much concerned with things having immediate psycholinguistic and/or cognitive-linguistic relevance, but it becomes just as clear that, to a considerable degree, it is linguists outside of CL that apply our methods, demonstrate their relevance to notions/data outside of the 'discourses', and validate some of the suggestions we've made. It follows that not only can CL benefit from relating to more of what happens in these 'irreconcilably different' disciplines, but that these disciplines have developed theories and models that allow us to move from the purely descriptive approach for which corpus linguists are often criticized to explanation, prediction, and the embedding into a larger context, or theory, or model. The kind of cognitive-linguistic/psycholinguistic model many of the above studies come with is an exemplar-based approach, which I need to briefly outline.

3. Exemplar-based models and their relation to CL

We have seen above that infants are very good at keeping track of distributional characteristics of the ambient language such as bigram probabilities (phonemes), phonological characteristics that help distinguish between open- and closed-class words or nouns and verbs, trigram probabilities (words), and phrasal boundaries as defined by function words. Obviously, distributional knowledge is ultimately knowledge based on frequencies of occurrence, frequencies of co-occurrence, and dispersion characteristics – but how is that acquired and represented? The main assumptions of exemplar-based models in linguistics are as follows.

Speakers/listeners encounter (aspects of) tokens/exemplars and 'place them' into a multidimensional memory space/network such that the location of that token is determined by the values it exhibits in the dimensions of the memory space. This is probably easiest to conceptualize for phonemes, because many of the dimensions with regard to which phonemes are described are inherently quantitative, such as their formant frequencies: a vowel sound will be placed at a location in the multidimensional space that corresponds to its perceived F_1, F_2, etc. frequencies. More generally, phonemes are "associated with a distribution of memory traces in a parametric space, in this case a cognitive representation of the parametric phonetic space" (Pierrehumbert 2003: 185).

Figure 1 represents a three-dimensional snapshot of a speaker's truly n>3-dimensional memory space, where percepts of some linguistic units are indicated as points whose locations are based on the three dimensions *x*, *y*, and *z*, and whose grey-shading reflects their positions on the *z*-axis. While the representation in merely three dimensions is of course a simplification, it is plain to see that, for example, the points make up two categories on the dimension represented on the *x*- and the *y*-axis: there is a category with low *x*- and high *y*-values and a category with high *x*- and low *y*-values. At the same time, these two categories appear to fall into two categories along the *z*-axis: low/dark values and high/light values. If the speaker whose memory system is represented in Figure 1 now perceives another linguistic unit of the type represented in Figure 1 – for example, a unit with the values x=0, y=12, z=-8, then a new (dark) point will be inserted at these coordinates and strengthen the representation of the category with low *x*- and *z*-values and high *y*-values. This also implies that exemplars which are similar/dissimilar to each other are in close proximity/at a distance from each other respectively, and categorization of a new exemplar proceeds on the basis of multidimensional spatial proximity to clouds of already memorized exemplars. In other words, "each instance redefines the system, however infinitesimally, maintaining its present state or shifting its probabilities in one direction or the other" (Halliday 2005: 67) and "it is usual that each learning event updates a statistical representation of a category independently of other learning events" (Ellis 2002: 147).

Two important qualifications are in order. First, the above does not imply that speakers/listeners remember each token and everything about each exemplar: while speakers do not immediately categorize tokens to discard all more detailed information, (aspects of) memories of individual exemplars may still not be accessible because they may

- decay or be subject to generalization/abstraction as well as reconstruction (Ellis 2002: 153; Abbot-Smith and Tomasello 2006: 275);
- never make it into long-term memory: the fact that "we normally don't remember things we encounter only once or twice (unless they are particularly striking, or highly significant for personal reasons)" (Dąbrowska 2009: 207) implicitly facilitates the identification of typical contexts.

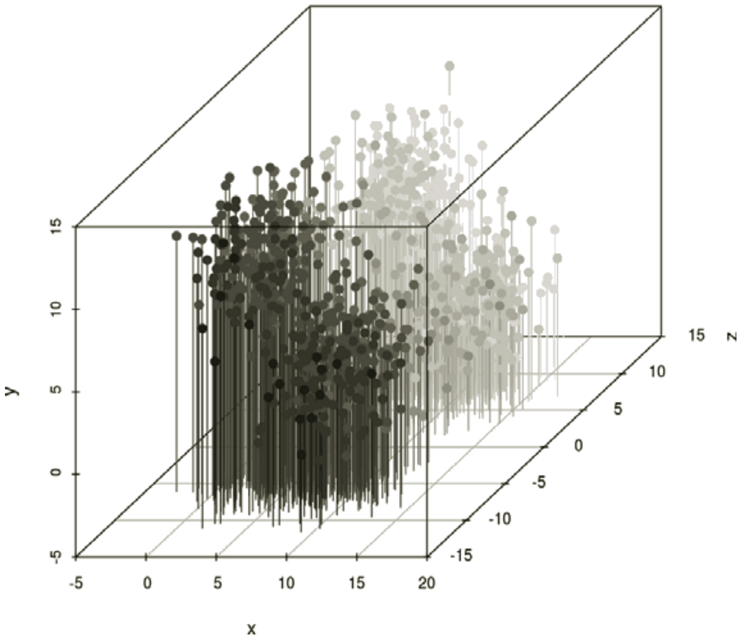

Figure 1. A three-dimensional representation of the *n*-dimensional memory space of a fictitious speaker of a language *L*

Second, this approach is not restricted to quantitative dimensions, such as formant frequencies. For example, in the case of words and constructions, it means that constructional slots are associated with distributions of words that occur in these slots and that in turn make up a (usually semantically fairly coherent) category. Even more generally, these kinds of effects are not restricted to phonemes or lexical items – quite the contrary: the distributional aspects to be remembered are numerous and involve phonetic, phonological, prosodic, morphemic, lexical (co-)occurrence as well as extra-linguistic/contextual aspects including utterance and situational context (such as the incongruity implication of the WXDY construction), sociolinguistic speaker factors, and information concerning register/mode, where I follow Biber (1995: 9) and understand register as a situationally/communicatively-defined category.

It may be useful to discuss the notion of register in this context a bit. Not only is register currently a hot topic in corpus linguistics, it is also an extremely complex and multifaceted notion and one that has repercussions for very many linguistic phenomena: depending on whom we talk to, we adjust our articulatory effort(s), lexical choices, syntactic complexity, etc. As Halliday (2005: 66) put it, "[r]egister variation can in fact be defined as systematic variation in probabilities; a register is a tendency to select certain combinations of meanings with certain frequencies". It seems counterintuitive, though, to assume that our cognitive systems have a different multidimensional space of the type schematically represented in Figure 1 for each register, especially since the number of registers we engage in, or are otherwise exposed to, is considerable. It is therefore worth mentioning at least briefly how something like register can be realistically integrated into this type of approach.

My take on this is that corpus-linguistic work contributes a very useful perspective on this. Biber's work on multidimensional variation provides the most instructive perspective and avoids the risk of postulating different knowledge/memory spaces for different registers. As is well known, Biber's approach to register variation is based on co-occurrence frequencies of many features from different linguistic levels of analysis; lexical, syntactic/ grammatical, semantic features, etc. Quite obviously, Biber's approach is therefore based on exactly the kind of information that is the foundation of exemplar-based models. Thus, the most straightforward way to include register information is in the form of the co-occurrence patterns of linguistic features with and within usage events in very much the same way that speakers store usage events with some contextual information (e.g., again, the incongruity implication of the WXDY construction), and in that case, what we call registers are clusters, or factors, that emerge from these co-occurrences and others. Alternatively, one could go a step further and postulate that the multidimensional space, in addition to the dimensions we already assume it to comprise (e.g. the three in Figure 1), also has dimensions such as the dimensions of variation identified in, say, Biber (1988).

The appealing aspects of this model, its implications, and the ways in which it is compatible with corpus work are manifold. On a theoretical level, it helps explain first language acquisition without recourse to largely untestable parameters, but also without a "meta-grammar" of the type mentioned by Maxwell (2010: 379) etc., a topic about which much of corpus linguistics proper has had little to say. It is compatible with our knowledge that speakers/listeners store immense amounts of probabilistic information, and the assumption of clouds of remembered exemplars can model all kinds of frequency effects:

- high frequencies of (co-)occurrence correspond to dense clouds with many different points in close proximity;
- categorization and prototype effects follow from the multidimensional structure of an exemplar cloud (e.g. exemplars in the 'middle' of a more densely populated area exhibit the prototype effect of first associations, faster recognition, etc.);
- the model can explain how, given their different exposure to language and, hence, different multidimensional spaces, even native speakers of a language can differ considerably in their command of the language and their judgments: "each speaker of a language has their own grammar, derived from the individual's linguistic experience. These grammars will obviously be similar and to a large extent overlapping, but they will not be identical". (Mason 2007: 2; cf. also Dąbrowska, submitted);
- the model can unproblematically account for register, sociolinguistic, and other contextual effects.

On a methodological level, this kind of model also forces us to turn (more) towards multidimensional approaches, which can be multidimensional in two senses: multidimensional$_1$ meaning that many different dimensions of variation are included in our analyses; multidimensional$_2$ meaning that co-occurrence information from as many different dimensions is included (cf. Gries 2010c). An example of an approach that is multidimensional$_1$ is the Behavioral Profile approach; examples for approaches that are multidimensional$_1$ and multidimensional$_2$ are association rules or multifactorial classification/ regression approaches, where model selection processes are used to determine which dimensions for which data are available should be retained (i.e., for which dimensions we need to rotate our multidimensional space to see another important difference). This implies the use of, for instance, more

- general(ized) linear models as in studies of alternation phenomena (Gries 2003; Szmrecsanyi 2005; Bresnan et al. 2007; Arppe 2008; Janda et al. 2010, etc.).
- more mixed-effects models (at least once these have been developed well enough), because they allow us to model all sorts of effects specific to speakers, writers, files, words, constructions, etc.
- more bottom-up and/or multivariate approaches in exploratory studies, in the parlance of an exemplar-model approach, to determine which (meaningful) dimensions emerge when the space is compressed and rotated; this includes principal component/factor analyses (Biber's multidimensional approach being the most fitting example); cluster analyses (as in Divjak and Gries 2006); correspondence analysis (as in Glynn 2008); multidimensional scaling (as in Croft and Poole 2008), etc.

4. Wrapping up

In this position paper, I tried to make a few minor proposals, which included the proposal to maybe rethink the contrast of corpus-driven and corpus-based linguistics, and to definitely rethink the us vs. them hijacking warfare. However, my main focus was something else: first, I hope I have been able to

- discuss some reasons why some part of theoretical linguists and some part of CL have so far not yet entered into the kind of fruitful relation I would like to see more of;
- convey my thoughts on why I think that this (only slowly narrowing) gap should be closed at a much faster pace;
- show that much of CL is extremely compatible with recent developments in cognitive linguistics/construction grammar as well as psycholinguistic approaches based on exemplar models, and that these theories can (i) help us answer *why*-questions in a much more revealing way than the humanistic hermeneutic-circle meaning-in-discourses-is-negotiated-by-the-community way upheld by some as well as (ii) inform corpus linguistics in terms of research questions we may want to ask next, methods from these areas that we can learn from a lot to improve the ways in which we pursue our questions, and findings that help us select which of the methods we have developed are most useful in which contexts. (Recent work especially in the domain of corpus-based SLA research is most instructive in this respect, cf., e.g., Ellis and Ferreira-Junior 2009).

Williams (2010: 402) claims that “Firth refuted all mentalism”, but, well, Firth maybe *argued against* mentalism, but certainly did not *refute* it, and I for one am not willing to settle for a corpus linguistics that tries to sell, or dignify, pointing to (repeated) co-occurrences in discourses as ‘explanation’. Is it not better to be able to explain distributions in corpora – e.g., reduced pronunciations of words – with reference to generally-known cognitive mechanisms regarding learning, habitualization, and articulatory routines than to what else happens in the discourse? Should we not explain repetitions of syntactic patterns based on implicit learning mechanisms that may ultimately lead to a unified approach towards syntactic and lexical priming and are attested in many other phenomena, too, rather than just point to repetition effects in discourses. Don’t we want to explain constructional choices such as Particle Placement with reference to cognitive mechanisms of online sentence production rather than just catalog which verbs prefer which construction? Is it not better 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 automatization as a result of frequency of occurrence than to what else happens in the discourse? I think the answer is always “Yes!”.

On that basis, my main focus is the proposal for us corpus linguists to assume as the main theoretical framework within which to explain and embed our analyses a psycholinguistically informed, (cognitively-inspired) exemplar/usage-based linguistics. Thankfully, I am not alone in this. There are some linguists who have assumed at least somewhat similar positions already (Schönefeld 1999; Schmid 2000; Mukherjee 2004; Butler 2004, for instance), but the major breakthrough I think is needed in order for corpus linguistics to shed its 'purely-descriptive' label has not yet happened. The from my point of view most important arguments in a very similar spirit are from Miller and Charles (e.g. 1991) as well as Hoey (e.g. 2005).

For example, Miller and Charles's work on near synonymy and antonymy (e.g. Miller and Charles 1991) involves the notion of a contextual representation, which is "a mental representation of the contexts in which the word occurs, a representation that includes all of the syntactic, semantic, pragmatic, and stylistic information required to use the word appropriately", but even more fitting is one of my favorite quotes from Hoey (2005: 11):

> the mind has a mental concordance of every word it has encountered, a concordance that has been richly glossed for social, physical, discoursal, generic and interpersonal context. […] [A]ll kinds of patterns, including collocational patterns, are available for use.

It's time to finally recognize this connection between corpus linguistics, cognitive linguistics and psycholinguistics…

Notes

[1] This paper is a revised and much extended version of Gries (2010a). In citations, I use "bcd" to refer to the bootcamp discourse on the CORPORA list in the summer 2008 that followed the announcement of my Quantitative Corpus Linguistics with R bootcamp (cf. <http://listserv.linguistlist.org/cgi-bin/wa?A1=ind0808&L=corpora> (14.09.2011)). Much of this article takes issue with Teubert's position, which is solely due to the facts that (i) he was for many years the editor of what probably is the flagship journal of the discipline, (ii) like Louw, he has been being very vocal with regard to his position, but (iii) unlike Louw, his position is comprehensible.

[2] This quote is from a 'review' considered anonymous of two book manuscripts submitted to the Benjamins SCL series in 2003/2004.

[3] Curiously, Teubert (bcd 2010: 354) even argues against a particular software that I have come to be associated with on the grounds that "it does not matter what kind of strings of information are processed. It could be language, but it could also be DNA sequences or the ciphers behind the '3' in the number pi" – as if that wasn't true of any concordance such as MicroConcord (which is used by Louw, but runs on DOS and seems unable to output more than 1500 matches (<http://www.lexically.net

/software/index.htm> (14.09.2011)) or uses non-ASCII characters (<http://www.athel.com/order/engsoft.html#mcc> (14.09.2011)).

4 I use the term *usage-based* here as meaning ‘linking use (as in ‘found in corpora’), synchrony, diachrony’ and in terms of Langacker’s (1987: 494) statement that “[s]ubstantial importance is given to the actual use of the linguistic system and a speaker’s knowledge of this use” (cf. Gonzálvez-García and Butler 2006 for more discussion).

References

Aarts, J. (2002), ‘Does corpus linguistics exist? Some old and new issues’, in: L.E. Breivik and A. Hasselgren (eds.), *From the COLT’s Mouth… and Others’: Language Corpora Studies in Honour of Anna-Brita Stenström.* Amsterdam: Rodopi. 1-19.

Abbot-Smith, K. and M. Tomasello (2006), ‘Exemplar-learning and schematization in a usage-based account of syntactic acquisition’, *The Linguistic Review*, 23: 275-90.

Albright, A. and B. Hayes (2003), ‘Rules vs. analogy in English past tenses: A computational/experimental study’, *Cognition*, 90: 119-61.

Ambridge, B., A. Theakston, E.V.M. Lieven and M. Tomasello (2006), ‘The distributed learning effect for children’s acquisition of an abstract grammatical construction’, *Cognitive Development*, 21: 174-93.

Arppe, A. (2008), *Univariate, Bivariate, and Multivariate Methods in Corpus-based Lexicography: A Study of Synonymy*. Unpublished Ph.D. dissertation, University of Helsinki.

Biber, D. (1988), *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1995), *Dimensions of Register Variation: A Cross-Linguistic Comparison.* Cambridge: Cambridge University Press.

Biber, D. (2009), ‘A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing’. Plenary talk held at Corpus Linguistics 2009, University of Liverpool.

Bowker, L. and J. Pearson (2002), *Working With Specialized Language: A Practical Guide to Using Corpora.* London: Routledge.

Bresnan, J., A. Cueni, T. Nikitina and R.H. Baayen (2007), ‘Predicting the dative alternation’, in: G. Bouma, I. Krämer and J. Zwarts (eds.), *Cognitive Foundations of Interpretation*. Amsterdam: Royal Netherlands Academy of Arts and Sciences. 69-94.

Butler, C.S. (2004), ‘Corpus studies and functional linguistic theories’, *Functions of Language*, 11: 147-86.

Bybee, J.L. (1985), *Morphology: A Study of the Relation Between Meaning and Form*. Amsterdam: John Benjamins.

Casenhiser, D.M. and A.E. Goldberg (2005), ‘Fast mapping of a phrasal form and meaning’, *Developmental Science*, 8: 500-508.

Croft, W. and K.T. Poole (2008), 'Inferring universals from grammatical variation: Multidimensional scaling for typological analysis', *Theoretical Linguistics*, 34: 1-37.

Croft, W. (2009), 'Toward a social cognitive linguistics', in: V. Evans and S. Pourcel (eds.), *New Directions in Cognitive Linguistics*. Amsterdam: John Benjamins. 395-420.

Dąbrowska, E. (2009), 'Words as constructions', in: V. Evans and S. Pourcel (eds.), *New Directions in Cognitive Linguistics*. Amsterdam: John Benjamins. 201-223.

Dąbrowska, E. (submitted), 'Individual differences in native language attainment: A review article'.

Divjak, D.S. and St.Th. Gries (2006), 'Ways of trying in Russian: Clustering behavioral profiles', *Corpus Linguistics and Linguistic Theory*, 2: 23-60.

Ellis, N.C. (2002), 'Frequency effects in language processing and acquisition', *Studies in Second Language Acquisition*, 24: 143-88.

Ellis, N.C. and F. Ferreira-Junior (2009), 'Constructions and their acquisition: Islands and the distinctiveness of their occupancy', *Annual Review of Cognitive Linguistics*, 7: 187-220.

Evert, S. and B. Krenn (2005), 'Using small random samples for the manual evaluation of statistical association measures', *Computer Speech and Language*, 19: 450-66.

Fidelholtz, J.L. (1975), 'Word frequency and vowel reduction in English', *Chicago Linguistic Society*, 11: 200-213.

Forster, K.I. and S.M. Chambers (1973), 'Lexical access and naming time', *Journal of Verbal Learning and Verbal Behavior*, 12: 627-635.

Glynn, D. (2008), 'Multivariate construction grammar: A quantitative approach to constructional semantics'. Paper presented at the conference of the German Cognitive Linguistics Association, University of Leipzig.

Goldberg, A.E. (2003), 'Constructions: A new theoretical approach to language', *Trends in Cognitive Sciences*, 7: 219-224.

Goldberg, A.E. (2006), *Constructions at Work: On the Nature of Generalization in Language*. Oxford: Oxford University Press.

Gonzálvez-García, F. and C.S. Butler (2006), 'Mapping functional-cognitive space', *Annual Review of Cognitive Linguistics*, 4: 39-96.

Gries, St.Th. (2003), *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London: Continuum.

Gries, St.Th. (2006a), 'Introduction', in: St.Th. Gries and A. Stefanowitsch (eds.), *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*. Berlin: de Gruyter. 1-17.

Gries, St.Th. (2006b), 'Exploring variability within and between corpora: Some methodological considerations', *Corpora*, 1:109-151.

Gries, St.Th. (2008), 'Dispersions and adjusted frequencies in corpora', *International Journal of Corpus Linguistics*, 13: 403-437.

Gries, St.Th. (2009), 'Dispersions and adjusted frequencies in corpora: Further explorations', in: St.Th. Gries, S. Wulff and M. Davies (eds.), *Corpus*

Linguistic Applications: Current Studies, New Directions. Amsterdam: Rodopi. 197-212.

Gries, St.Th. (2010a), 'Corpus linguistics and theoretical linguistics: A love-hate relationship? Not necessarily…', *International Journal of Corpus Linguistics*, 15: 327-343.

Gries, St.Th. (2010b), 'Bigrams in registers, domains, and varieties: A bigram gravity approach to the homogeneity of corpora', in: M. Mahlberg, V. González-Diaz and C. Smith (eds.), *Proceedings of Corpus Linguistics 2009*, University of Liverpool. <http://ucrel.lancs.ac.uk/publications /cl2009/404_FullPaper.doc> (14.09.2011).

Gries, St.Th. (2010c), 'Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics', *The Mental Lexicon*, 5: 323-346.

Gries, St.Th. (forthcoming), 'Methodological and interdisciplinary stance in corpus linguistics', in: G. Barnbrook, V. Viana and S. Zyngier (eds.), *Perspectives on Corpus Linguistics: Connections and Controversies*. Amsterdam: John Benjamins.

Gries, St.Th., B. Hampe and D. Schönefeld (2005), 'Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions', *Cognitive Linguistics*, 16: 635-676.

Gries, St.Th., B. Hampe and D. Schönefeld (2010), 'Converging evidence II: More on the association of verbs and constructions', in: J. Newman and S. Rice (eds.), *Empirical and Experimental Methods in Cognitive/Functional Research*. Stanford, CA: CSLI. 59-72.

Gries, St.Th. and J. Mukherjee (2010), 'Lexical gravity across varieties of English: An ICE-based study of *n*-grams in Asian Englishes', *International Journal of Corpus Linguistics*, 15: 520-548.

Gries, St.Th., J. Newman and C. Shaoul (2011), '*N*-grams and the clustering of genres', *Empirical Language Research*, 5. <http://ejournals.org.uk/ELR/ article/2011/1> (10.11.2011).

Halliday, M.A.K. (2005), *Computational and Quantitative Studies*. London: Continuum.

Hardie, A. and T. McEnery (2010), 'On two traditions in corpus linguistics, and what they have in common', *International Journal of Corpus Linguistics*, 15: 384-394.

Hawkins, J.A. (1994), *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Hawkins, J.A. (2005), *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Hoey, M. (2005), *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Howes, D.H. and R.L. Solomon (1951), 'Visual duration threshold as a function of word probability', *Journal of Experimental Psychology*, 41: 401-410.

Hunston, S. and G. Francis (2000), *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Janda, L., T. Nesset and R.H. Baayen (2010), 'Capturing correlational structure in Russian paradigms: A case study in logistic mixed-effects modeling', *Corpus Linguistics and Linguistic Theory*, 6: 29-48.

Kita, K., Y. Kato, T. Omoto and Y. Yano (1994), 'A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria', *Journal of Natural Language Processing*, 1: 21-33.

Krug, M. (1998): 'String frequency: A cognitive motivating factor in coalescence, language processing, and linguistic change', *Journal of English Linguistics*, 26: 286-320.

Langacker, R.W. (1987), *Foundations of Cognitive Grammar. Volume I: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

Leech, G.N. (1992), 'Corpora and theories of linguistic performance', in: J. Svartvik (ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82*. Berlin: de Gruyter. 105-122.

Lyne, A.A. (1985), *The Vocabulary of French Business Correspondence*. Geneva: Slatkine-Champion. 101-124.

MacDonald, M.C. (1993), 'The interaction of lexical and syntactic ambiguity', *Journal of Memory and Language*, 32: 692-715.

Mason, O. (2006), *The Automatic Extraction of Linguistic Information from Text Corpora*. Unpublished Ph.D. dissertation, University of Birmingham.

Mason, O. (2007), 'From lexis to syntax: The use of multi-word units in grammatical description', in: C. Camuali, M. Constant and A. Dister (eds.), *Proceedings of Lexis and Grammar 2007*. <http://infolingu.univ-mlv.fr/Colloques/Bonifacio/proceedings/mason.pdf> (14.09.2011).

Maxwell, M. (2010), 'Limitations of corpora', *International Journal of Corpus Linguistics*, 15: 379-383.

McDonald, S., R. Shillcock and C. Brew (2001), 'Low-level predictive inference in reading: Using distributional statistics to predict eye movements'. Paper presented at the 7th Annual Conference on Architectures and Mechanisms for Language Processing, Saarbrücken.

McEnery, T. and A. Wilson (1996), *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T., R. Xiao and Y. Tono (2006), *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

Meyer, C.F. (2002), *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Miller, G.A. and W.G. Charles (1991), 'Contextual correlates of semantic similarity', *Language and Cognitive Processes*, 6: 1-28.

Mintz, T.H., E.L. Newport and T.G. Bever (2002), 'The distributional structure of grammatical categories in the speech to young children', *Cognitive Science*, 26: 393-424.

Mukherjee, J. (2004), 'Corpus data in a usage-based cognitive grammar', in: K. Aijmer and B. Altenberg (eds.), *Corpus Data in a Usage-based Cognitive Grammar*. Amsterdam: Rodopi. 85-100.

Mukherjee, J. (2010), 'Corpus linguistics versus corpus dogmatism – pace Wolfgang Teubert', *International Journal of Corpus Linguistics*, 15: 370-378.

Nagao, M. and S. Mori. (1994), 'A new method of *n*-gram statistics for large number of *n* and automatic extraction of words and phrases from large text data of Japanese', *Proceedings of the 15th Conference on Computational Linguistics*. 611-615. <http://dl.acm.org/ft_gateway.cfm?id=991994&type=pdf&CFID=45628901&CFTOKEN=82898916> (14.09.2011).

Pierrehumbert, J. (2003), 'Probabilistic phonology: Discrimination and robustness', in: R. Bod, J. Hay, and S. Jannedy (eds.), *Probabilistic Linguistics*. Cambridge, MA: The M.I.T. Press. 177-228.

Redington, M., N. Chater and S. Finch (1998), 'Distributional information: A powerful cue for acquiring syntactic categories', *Cognitive Science*, 22: 435-469.

Saffran, J.R., R.N. Aslin and E.L. Newport (1996), 'Statistical learning by 8-months-old infants', *Science*, 274: 1926-1928.

Schmid, H.-J. (2000), *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Berlin: de Gruyter.

Schmidtke-Bode, K. (2009), '*Going-to*-V and *gonna*-V in child language: A quantitative approach to constructional development', *Cognitive Linguistics*, 20: 509-538.

Schönefeld, D. (1999), 'Corpus linguistics and cognitivism', *International Journal of Corpus Linguistics*, 4: 131-171.

Simpson, R. and N.C. Ellis (2005), 'An academic formulas list: Extraction, validation, prioritization'. Paper presented at Phraseology 2005, Université catholique de Louvain.

Sinclair, J.M. (1991), *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J.M. (2006), *Linear Unit Grammar: Integrating Speech and Writing*. Amsterdam: John Benjamins.

Snider, N. and I. Arnon (forthcoming), 'More than words: Speakers are sensitive to the frequency of multi-word sequences'.

Stubbs, M. (1993), 'British traditions in text analysis: From Firth to Sinclair', in: M. Baker, F. Francis and E. Tognini-Bonelli (eds.), *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins. 1-46.

Szmrecsanyi, B. (2005), 'Language users as creatures of habit: A corpus-based analysis of persistence in spoken English', *Corpus Linguistics and Linguistic Theory*, 1: 113-150.

Taylor, C. (2008), 'What is corpus linguistics? What the data says', in: M. Kytö and A.-B. Stenström (eds.), *ICAME Journal* 32: 179-200.

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

Teubert, W. (2010), 'Our brave new world?', *International Journal of Corpus Linguisitcs*, 15: 354-358

Tognini-Bonelli, E. (2001), *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Wiechmann, D. (2008), 'On the computation of collostruction strength: Testing measures of association as expressions of lexical bias', *Corpus Linguistics and Linguistic Theory*, 4: 253-290.

Williams, G. (2006), 'La linguistique de corpus: Une affaire prépositionnelle', in: F. Rastier and M. Ballabriga (eds.), *Corpus en Lettres et Sciences Sociales: Des Documents Numériques à l'Interprétation*. Paris: Texto. 151-158.

Williams, G. (2010), 'Many rooms with corpora', *International Journal of Corpus Linguistics*, 15: 400-407.

Wynne, M. (2010): 'Interdisciplinary relationships', *International Journal of Corpus Linguistics*, 15: 425-427.

Xiao, R. (2009), 'Theory-driven corpus research: Using corpora to inform aspect theory', in: A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics: An International Handbook*. Berlin: de Gruyter. 987-1008.