Corpus linguistics and theoretical linguistics

A love–hate relationship? Not necessarily…*

Stefan Th. Gries
University of California, Santa Barbara

> *[I]t is common now to address theoretical issues through the examination of bodies of naturally occurring language use.*
>
> (Bybee 2006: 712)

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 two 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.

1.1 Within CL: What we think corpus linguistics is

As for the former, some consider CL a theory, for instance Leech (1992: 106), Stubbs (1993: 2f.), Tognini-Bonelli (2001: 1), Teubert (2005: 2).[1] Others consider CL a methodology, such as McEnery & Wilson (1996), Meyer (2002), Bowker & Pearson (2002), McEnery et al. (2006: 7f.), Hardie & McEnery (this volume). 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 et al. 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[2])

International Journal of Corpus Linguistics 15:3 (2010), **327–343**. DOI 10.1075/ijcl.15.3.02gri
ISSN 1384–6655 / E-ISSN 1569–9811

Yet other corpus linguists (e.g. Aarts 2002, Teubert 2005, Williams 2006) use 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).

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).

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

> 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 linguists approach corpus data with moderate corpus-external premises, with the aim of testing and improving such theories, and often use corpus annotation. I myself am more of a corpus-based linguist and consider CL "a major methodological paradigm in applied and theoretical linguistics" (Gries 2006: 191). Why? First, I agree with Aarts, who coined the term *corpus linguistics*, but is also

> 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 theory called *experimental linguistics* or *self-paced reading time linguistics* even though, just like corpus-based results, 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 Linear Unit Grammar, I have yet to see a truly corpus-driven approach. Even self-proclaimed corpus-driven studies are often not as corpus-driven as they could/claim to be, which can be seen on different levels:

i. From a theoretical perspective, a truly corpus-driven approach would, strictly speaking, require a complete distributional analysis of the corpus (with maybe some machine-learning algorithm/neural network) to initially identify the linguistic units manifested in the data (as in, say, C.C. Fries's 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."[3]),

– many corpus-driven studies start out from some notion of a word;
– Bill Louw (2000) shows a concordance of *all sorts of*, and my guess is he has not first used a bottom-up or even replicable algorithm to ensure that *all sorts of* is in fact a unit;
– 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 & 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".

ii. From a more concrete perspective, the corpus-driven approach is also not always really corpus-driven. 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 *n* is indeed the best number for all *n*-grams; as one of few exceptions, Biber (2009) checks for 5-grams;
– check whether it would not indeed be better to have different *n*'s for different *n*-grams (cf. Kita et al. 1994, Mason 2006, and Mukherjee & Gries 2009 on varying length *n*-grams);
– even consider discontinuous *n*-grams (e.g. Nagao & Mori 1994).

iii. From a register perspective: Halliday (2005: 66) wrote "Register 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". I agree, but with regard to much corpus-driven work the register distinctions in a corpus may not be most useful from a truly bottom-up perspective (cf. Gries forthcoming, Gries et al. 2009).

Comparing corpus-driven and 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 seems a myth at best. 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 CL: 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 several things that make corpus linguistics less attractive to the observer from theoretical linguistics (apart from very time-consuming retrieval and coding operations). These include some corpus linguists'

i. rather "unusual" ideas about potentially relevant neighboring disciplines;
ii. rather "unusual" ways of defending their perspective(s);
iii. rather "unusual" ideas about the nature of the discipline (above and beyond the above issues).

I cannot discuss all these issues in detail so some examples must suffice. As for (i),

– Teubert (this volume) characterizes the relation between NLP and cognitive linguistics (CogLing) as follows: NLP is one of CogLing's "illegitimate offspring[s]". This just leaves me baffled. 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*.
– In a generally interesting paper, Mason (2007: 2) argues in favor of Linear Unit Grammar:

> "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".
>
> This betrays a serious misunderstanding of psycholinguistic approaches to language production and comprehension: An incremental approach to language production and comprehension does *by no means* require abandoning a largely hierarchical view of language structure (cf. Hawkins's 1994, 2005 or G. Kempen's work). 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),

- I find that, for instance, Fillmore's polarization of arm-chair linguists vs. corpus linguists is too often taken way too seriously: theoretical linguists need (corpus) data, but CL also needs (more) theory.
- Discourse with and about (more) theoretical linguistics is often characterized by a strange us vs. them gate-keeping warfare that (i) uses geographical labels in place of arguments (as when agendas are simply labeled as "transatlantic"), that (ii) contrasts (good) old-fashioned Sinclairian core corpus linguistics with those who "piss into"[4] Sinclair's canonical corpus linguistics tent and use corpora in "a seemingly inappropriate, toolbox-like, inherently non-Sinclairian way" (Mukherjee,[5] characterizing the viewpoint he opposes), that (iii) couches interdisciplinary discourse in terms of "hijacking" (Teubert, this volume) and "takeover bid[s]" (Williams, this volume).[6]

As for (iii),

- "corpus linguistics looks at phenomena which cannot be explained by recourse to general rules and assumptions" (Teubert 2005: 5) — I know many corpus linguists who are interested in explaining phenomena this way, especially since "general rules and assumptions" does 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) — 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 "[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 (2005: 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.

I don't see how blanking out the very thing that makes us human — mind/*Geist* — helps in the endeavor to make sense of the human condition…

1.3 Taking stock…

None of this must distract from the fact that CL 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 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 of measures of collocational strength, but not only do few corpus linguists try to determine which are best when (Evert & 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) find $p_{\text{Fisher-Yates}}$ to be best; Wiechmann (2008) gets the best results with minimum sensitivity, etc. 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 linguistics as a whole and in disciplines that have often independently arrived at similar conclusions, but also because external validation would streamline corpus-linguistic research. Butler (2004) argues that, contrary to what Sinclair believed, CL should not and need not be non-cognitive. A similar plea for the coming together of CL and cognitive approaches is made by Hoey (2005: 7), who states the need for a "greater awareness in corpus linguistics of the need for a more powerful and cognitively valid theory". 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. 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.[7] And in some sense, that is the logical choice. First, as we're talking about the *humanistic* perspective and the Geistes*wissenschaften*, isn't 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? Again, how can we seriously be in the *Geisteswissenschaften* if the one thing we *a priori* blank out is *Geist*?!

Second, at some point of 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 that input is also determined by that hearer's internal structure. As Maxwell put it:[8]

> 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[9] for a similar statement)

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

> *At a recent conference devoted to modern developments in corpus studies I was struck by the way that a number of speakers at the conference were setting up an opposition between 'corpus linguistics' and 'theoretical linguistics' — not a conflict, I mean, but a distinction, as if these were members of two distinct species. I commented on this at the time, saying that I found it strange because corpus linguistics seemed to me to be, potentially at least, a highly theoretical pursuit.*
>
> (Halliday 2005: 130)

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

2.1 CL and cognitive linguistics/psycholinguistics: Some commonalities

If one takes a look at some such frameworks (cf. Gonzálvez-García & Butler 2006 for excellent discussion), many commonalities between CL and (newer) developments in psycholinguistically informed, (cognitively-inspired) usage-based linguistics emerge. In fact, many things in CL have immediate psycholinguistic and/or cognitive-linguistic relevance:

i. When corpus linguists talk about *token frequencies*,
   - cognitive linguists become interested because, on the whole, token frequencies correlate with degree of entrenchment (Schmid 2000); with phonetic reduction and development of new forms (Fidelholtz 1975, Bybee & Thompson 1997, Bybee & Scheibman 1999); with resistance to morphosyntactic language change (Bybee & Thompson 1997), etc.;
   - psycholinguists become interested because, on the whole, token frequencies correlate with ease/earliness of acquisition (Casenhiser & Goldberg 2005); with lexical decision tasks, word naming, picture naming (Howes & Solomon 1951, Forster & Chambers 1973; re web data, cf. Van Durme et al., in progress).

ii. When corpus linguists 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 & Hayes 2003);
   - psycholinguists become interested because type frequencies are correlated with the productivity of, say, constructions in first/second language acquisition.

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

iv. 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.

v. 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 linguistics become interested because it reminds them of
   - Langacker's (1987: 57) 'unit', "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".
   - Langacker's (2000: 2) 'rule-list fallacy': "[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".

vi. 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 patterns are very similar to Goldberg's constructions:

> 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 & 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)

vii. 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 & Arnon, in progress);

In addition, 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 an ironic twist, many of these approaches are therefore more corpus-driven than much self-proclaimed corpus-driven work, as when Reddington et al. (1998) and Mintz et al. (2002) apply nearly completely 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 are now arguing for a cognitive sociolinguistics, and 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 corpus-related topics at several recent CogLing 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 …

Obviously, CL is concerned with many things having immediate psycholinguistic and/or cognitive-linguistic relevance, but it becomes just as clear that it is often linguists *outside of* CL that apply our methods, demonstrate their relevance to notions/data outside of the 'discourses', and validate 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 CL is 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, in which

> each instance redefines the system, however infinitesimally, maintaining its present state or shifting its probabilities in one direction or the other
>
> (Halliday 2005: 67)

> each learning event updates a statistical representation of a category independently of other learning events.
>
> (Ellis 2002: 147)

Speakers/listeners remember (aspects of) tokens/exemplars and "place them" into a multidimensional space/network (cf. Pierrehumbert 2003). As indicated above, the distributional aspects to be remembered are various and involve phonetic, phonological, prosodic, morphemic, lexical (co-)occurrence and extra-linguistic/contextual aspects including utterance context (e.g. the incongruity implication of the *What's*-X-*Doing*-Y construction), sociolinguistic speaker factors, and information about register/genre/mode.

Exemplars which are identical to an already memorized exemplar (at the available level of granularity) strengthen the existing exemplar's representation. Exemplars which are similar/dissimilar to each other are close to/far from each other respectively, and categorization of a new exemplar proceeds on the basis of multidimensional spatial proximity to clouds of already memorized exemplars. This does not mean that speakers/listeners remember each exemplar and everything about it: (aspects of) memories of individual exemplars may not be accessible because they may decay, be generalized over, or never make it into long-term memory.

The appealing aspects of this model, its implications, and its compatibility with CL are manifold: it explains first language acquisition without recourse to largely untestable parameters etc., a topic about which much of CL proper has had little to say. It "embodies" our knowledge that speakers/listeners store immense amounts of probabilistic information, and the assumption of clouds of remembered exemplars models all kinds of frequency effects: high frequencies of (co-) occurrence correspond to particularly dense clouds with many different points in close proximity (from different angles). That in turn means that this kind of approach can easily account for categorization/prototype effects, differences between native speakers of the same language, and register or other contextual effects, which merely become additional dimensions of variation.

> *there is a major convergence of research from many different perspectives — corpus-based analysis, computational linguistics, discourse, cognitive and functional linguistics, and psycholinguistics — that all point to a new theory of grammar with its attendant theory of language acquisition.*
>
> (Bybee 2002: 215)

3. Wrapping up

I made 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 wanted to

- discuss reasons why some part of theoretical linguistics and some part of CL have so far not yet entered into the kind of fruitful relation I would like to see more;
- explain why I think that this gap should be closed at a much faster pace;
- show that much of CL is extremely compatible with developments in CogLing/Construction Grammar and with some psycholinguistic theories/models, and that these theories can help CL answer *why*-questions in a better way than the humanistic hermeneutic-circle meaning-in-discourses-is-negotiated-by-the-community way upheld by some.

Thus, my main proposal is for us corpus linguists to assume as the main theoretical framework within which to explain and embed our analyses a psycholinguistically informed, (cognitively-inspired) usage-based linguistics. Thankfully, I am not alone in this. There are some linguists who have assumed similar positions already (Schönefeld 1999, Schmid 2000, Mukherjee 2004, and Butler 2004, for instance), but the major breakthrough I would hope for has not yet happened. The from my point of view most important arguments in a very similar spirit are from Miller and Charles as well as Hoey. For example, Miller and Charles's work on near synonymy and antonymy (e.g. Miller & Charles 1991: 26) involves the notion of a *contextual representation*, "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. […] all 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

* I thank Chris Butler for comments on an earlier draft. The usual disclaimers apply.

1. Much of this article takes issue with Teubert's position, which is due to his having been the editor of what probably is the flagship journal of the discipline and his being very vocal with regard to his position.

2. Hardie, message # 12240 to Corpora List, 14 August 2008; see also Hardie & McEnery (this volume) for a more extended discussion.

3. Teubert, message # 12237 to Corpora List, 14 August 2008; see also Teubert (this volume) and Mukherjee (this volume).

4. This quote is from a "review" considered anonymous of two book manuscripts submitted to a book series in 2003/2004.

5. Mukherjee, message # 12250 to Corpora List, 16 August 2008; see also Mukherjee (this volume) for a more extended discussion.

6. Teubert (this volume) 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 bit 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 concordancer …

7. 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 & Butler 2006 for more discussion).

8. Maxwell, message # 12261 to Corpora List, 17 August 2008; see also Maxwell (this volume) for a more extended discussion.

9. Washtell, message # 12294 to Corpora List, 21 August 2008.

References

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

Albright, A. & Hayes, B. 2003. "Rules vs. analogy in English past tenses: A computational/experimental study". *Cognition*, 90 (2), 119–161.

Ambridge, B., Theakston, A., Lieven, E. V. M. & Tomasello, M. 2006. "The distributed learning effect for children's acquisition of an abstract grammatical construction". *Cognitive Development*, 21 (2), 174–193.

Biber, D. 2009. "A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing". *International Journal of Corpus Linguistics*, 14 (3), 275–311.

Bowker, L. & Pearson, J. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.

Butler, C. S. 2004. "Corpus studies and functional linguistic theories". *Functions of Language*, 11 (2), 147–186.

Bybee, J. L. 1985. *Morphology: A Study of the Relation between Meaning and Form*. Amsterdam/Philadelphia: John Benjamins.

Bybee, J. L. 2002. "Phonological evidence for exemplar storage of multiword sequences". *Studies in Second Language Acquisition*, 24 (2), 215–221.

Bybee, J. L. 2006. "From usage to grammar: The mind's response to repetition". *Language*, 82 (4), 711–733.

Bybee, J. L. & Scheibman, J. 1999. "The effect of usage on degrees of constituency: The reduction of *don't* in English". *Linguistics*, 37 (4), 575–596.

Bybee, J. L. & Thompson, S. A. 1997. "Three frequency effects in syntax". *Berkeley Linguistics Society*, 23, 65–85.

Casenhiser, D. M. & Goldberg, A. E. 2005. "Fast mapping of a phrasal form and meaning". *Developmental Science*, 8 (6), 500–508.

Croft, W. 2009. "Toward a social cognitive linguistics". In V. Evans & S. Pourcel (Eds.), *New Directions in Cognitive Linguistics*. Amsterdam/Philadelphia: John Benjamins, 395–420.

Ellis, N. C. 2002. "Frequency effects in language processing and acquisition". *Studies in Second Language Acquisition*, 24 (2), 143–188.

Evert, S. & Krenn, B. 2005. "Using small random samples for the manual evaluation of statistical association measures". *Computer Speech and Language*, 19 (4), 450–466.

Fidelholtz, J. L. 1975. "Word frequency and vowel reduction in English". *Chicago Linguistic Society*, 11, 200–213.

Forster, K. I. & Chambers, S. M. 1973. "Lexical access and naming time". *Journal of Verbal Learning and Verbal Behavior*, 12 (6), 627–635.

Goldberg, A. E. 2003. "Constructions: A new theoretical approach to language". *Trends in Cognitive Sciences*, 7 (5), 219–224.

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

Gonzálvez-García, F. & Butler, C. S. 2006. "Mapping functional-cognitive space". *Annual Review of Cognitive Linguistics*, 4, 39–96.

Gries, St. Th. 2006. "Introduction". In St. Th. Gries & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis*. Berlin/New York: Mouton de Gruyter, 1–17.

Gries, St. Th. 2008. "Dispersions and adjusted frequencies in corpora". *International Journal of Corpus Linguistics*, 13 (4), 403–437.

Gries, St. Th. 2009. "Dispersions and adjusted frequencies in corpora: Further explorations". In St. Th. Gries, S. Wulff, & M. Davies (Eds.), *Corpus Linguistic Applications: Current Studies, New Directions*. Amsterdam: Rodopi, 197–212.

Gries, St. Th. Forthcoming. "Bigrams in registers, domains, and varieties: A bigram gravity approach to the homogeneity of corpora". *Proceedings of Corpus Linguistics Conference 2009*, University of Liverpool.

Gries, St. Th., Hampe, B. & Schönefeld, D. 2005. "Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions". *Cognitive Linguistics*, 16 (4), 635–676.

Gries, St. Th., Newman, J. & Shaoul, C. 2009. "*N*-grams and the clustering of genres". Paper presented at the workshop "Corpus, colligation, register variation". *31st Annual Meeting of the Deutsche Gesellschaft für Sprachwissenschaft, University of Osnabrück, 4–6 March*.

Halliday, M. A. K. 2005. *Computational and Quantitative Studies*. London/New York: Continuum.

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/New York: Routledge.

Howes, D. H. & Solomon, R. L. 1951. "Visual duration threshold as a function of word probability". *Journal of Experimental Psychology*, 41 (6), 401–410.

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

Kita, K., Kato, Y., Omoto, T. & Yano, Y. 1994. "A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria". *Journal of Natural Language Processing*, 1 (1), 21–33.

Krug, M. 1998. "String frequency: A cognitive motivating factor in coalescence, language processing, and linguistic change". *Journal of English Linguistics*, 26 (4), 286–320.

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

Langacker, R. W. 2000. "A dynamic usage-based model". In M. Barlow & S. Kemmer (Eds.), *Usage-based Models of Language*. Stanford: CSLI Publications, 1–63.

Leech, G. N. 1992. "Corpora and theories of linguistic performance". In Jan Svartvik (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4–8 August*. Berlin/New York: Mouton de Gruyter, 105–122.

Louw, W. 2000. "Contextual prosodic theory: Bringing semantic prosodies to life". In C. Heffer & H. Sauntson (Eds.), *Words in Context, a Tribute to John Sinclair on his Retirement*. Birmingham: ELR.

Lyne, A. A. 1985. *The Vocabulary of French Business Correspondence*. Geneva/Paris: Slatkine-Champion.

MacDonald, M. C. 1993. "The interaction of lexical and syntactic ambiguity". *Journal of Memory and Language*, 32 (5), 692–715.

Mason, O. 2006. "The automatic extraction of linguistic information from text corpora". Unpublished PhD dissertation, University of Birmingham.

Mason, O. 2007. "From lexis to syntax: The use of multi-word units in grammatical description". *Proceedings of Lexis and Grammar 2007, Bonifacio, Corsica, 2–6 October*. Available at: http://infolingu.univ-mlv.fr/english/Colloques/Bonifacio/proceedings.html (accessed May 2010).

McDonald, S., Shillcock, R. & Brew, C. 2001. "Low-level predictive inference in reading: Using distributional statistics to predict eye movements". *7th Annual Conference on Architectures and Mechanisms for Language Processing, Saarbrücken, Germany, 20–22 September*. Available at: http://www.amlap.org/2001/proc/proceedings_online.html (accessed May 2010).

McEnery, T. & Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T., Xiao, R. & Tono, Y. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London/New York: Routledge.

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

Miller, G. A. & Charles, W. G. 1991. "Contextual correlates of semantic similarity". *Language and Cognitive Processes*, 6 (1), 1–28.

Mintz, T. H., Newport, E. L. & Bever, T. G. 2002. "The distributional structure of grammatical categories in the speech to young children". *Cognitive Science*, 26 (4), 393–424.

Mukherjee, J. 2004. "Corpus data in a usage-based cognitive grammar". In K. Aijmer & B. Altenberg (Eds.), *Corpus Data in a Usage-based Cognitive Grammar*. Amsterdam: Rodopi, 85–100.

Mukherjee, J. & Gries, St. Th. 2009. "Lexical gravity across varieties of English: An ICE-based study of speech and writing in Asian Englishes". Paper presented at *ICAME 2009, University of Lancaster, 27–31 May*.

Nagao, M. & Mori, S. 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*, Vol. 1, 611–615.

Pierrehumbert, J. 2003. "Probabilistic phonology: Discrimination and robustness". In R. Bod, J. Hay & S. Jannedy (Eds.), *Probabilistic Linguistics*. Cambridge, MA: The MIT Press, 177–228.

Reddington, M., Chater, N. & Finch, S. 1998. "Distributional information: A powerful cue for acquiring syntactic categories". *Cognitive Science*, 22 (4), 435–469.

Saffran, J. R., Aslin, R. N. & Newport, E. L. 1996. "Statistical learning by 8-month-old infants". *Science*, 274 (5294), 1926–1928.

Schmid, H.-J. 2000. *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Berlin/New York: Mouton de Gruyter.

Schmidtke-Bode, K. 2009. "*Going-to*-V and *gonna-V* in child language: A quantitative approach to constructional development". *Cognitive Linguistics*, 20 (3), 509–538.

Schönefeld, D. 1999. "Corpus linguistics and cognitivism". *International Journal of Corpus Linguistics*, 4 (1), 131–171.

Simpson, R. & Ellis, N. C. 2005. "An academic formulas list: Extraction, validation, prioritization". Paper presented at *Phraseology 2005, Université catholique de Louvain, 13–15 October*.

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

Snider, N. & Arnon, I. In progress. "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 & E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins, 1–46.

Taylor, C. 2008. "What is corpus linguistics? What the data says". *ICAME Journal*, 32, 179–200.

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

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins.
Van Durme, B., Austin, F. & Jaeger, T. F. In progress. "Resources for cross-linguistic psycholinguistic research: The web is good enough".
Wiechmann, D. 2008. "On the computation of collostruction strength: Testing measures of association as expressions of lexical bias". *Corpus Linguistics and Linguistic Theory*, 4 (2), 253–290.
Williams, G. 2006. "La linguistique de corpus: Une affaire préposition- nelle". In F. Rastier & M. Ballabriga (Eds.), *Corpus en Lettres et Sciences Sociales: Des Documents Numériques à l'Interprétation*. Paris: Texto, 151–158.
Xiao, R. 2009. "Theory-driven corpus research: Using corpora to inform aspect theory". In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook*. Berlin/New York: Mouton de Gruyter, 987–1008.

Author's address

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
Santa Barbara, CA 93106-3100
United States of America

stgries@linguistics.ucsb.edu