AIMING LOW IN LINGUISTICS: LOW-LEVEL GENERALIZATIONS IN CORPUS-BASED RESEARCH

John Newman
University of Alberta

Corpora have played an important role in modern linguistics. I review some of the ways in which corpora have been relied upon in linguistics and how they have become increasingly common as sources of data in linguistic research. I then illustrate how corpora allow linguists to explore low-level patterns of co-occurrence associated with the verb in English. The corpus-based research reported here illustrates just one aspect of a larger empirical turn in recent linguistics which views corpus-based and experimentally based linguistic research as crucial, even foundational, in the future evolution of the field, correcting a severe imbalance in the recent history of linguistics which has favored native speaker intution as the source of linguistic evidence.

1. Introduction

A corpus-based approach to the study of language allows linguists to investigate language patterns at the level of word form which, in the case of an inflected language, means at the level of inflected word form. For the most part, linguists have preferred to work at a rather more abstract level of language where it is the lemma (an abstraction over related inflectional forms of a paradigm) which is the focus of attention, rather than the inflected word forms. The patterns which can be observed when we choose to focus on word forms and their contexts of usage are relatively "low-level", comparted with patterns based on lemmas. I believe that there is much to be learned from investigating these low-level patterns in language and illustrate this approach below, in Section 3. A corpus-based approach which explores low-level generalizations in the use of language invariably yields a very large number of observations which must be reconciled with other kinds of empirical and theoretical ideas about language if we are to make progress in linguistics and in Section 4 I offer my own views about how we might proceed to construct a broad, empirically informed theory of language. I begin with a brief overview of corpus-based research in the recent history of linguistics.

2. Corpus-based research

2.1 The invisible corpora of modern linguistics

Corpora, understood as collections of written texts or transcribed spoken discourse, have always played a role in modern linguistics, even if that role has not always been very explicitly acknowledged or emphasized. Where we have earlier stages of a language attested in written form, for example, those earlier texts are inevitably the basis for linguistic research. Studying such texts is, of course, not the only way to carry out historical research on a language. The comparative method of historical reconstruction provides ways to reconstruct some features of earlier stages of languages, whether or not a textual record exists. But when it comes to studying earlier stages of language above the word level, such as phrasing, word order, clause structure, rhetorical style etc., then earlier texts are inevitably the primary source of evidence. For some languages, such as Gothic, the corpus represents the complete record of the language, unlike the usual situation surrounding corpora of living languages where corpora are, at best, representative, but by no means exhaustive, of language use. One has no choice but to work with the Gothic corpus if one is researching Gothic syntax or semantics. The reliance on corpus data is so obvious in such cases that it would be unusual for specialists in Gothic to announce that they are "corpus linguists", even though they could be properly described in this way. The same applies to anyone working on stylistic or other linguistic analysis of the literature of an author or a specific period.

The linguistic study of first language acquisition has long made extensive use of diaries and journals of children's speech and it is not surprising that it is in the domain of child language studies that some of the first spoken corpora and corpus tools were first developed. The CHILDES collection of child language corpora (http://childes.psy.cmu.edu/) has grown to be an invaluable source of data for such research, while the CLAN tool,

developed for use with such corpora, continues to be one of the most useful tools for retrieving patterns and frequencies from child corpora as well as conversational corpora more generally. A modern and extreme example of a language acquisition corpus is the Human Speechome Project, led by Deb Roy, Director of the M.I.T. Media Lab, and described further at http://www.media.mit.edu/cogmac/projects/hsp.html. Roy, beginning in 2005, has been recording his son's language development within the home of the parents and the caregivers by gathering approximately 10 hours of audio and video on a daily basis from birth to age three. In 2006 already, the corpus contained 24,000 hours of video and 33,000 hours of audio recordings (Roy et al. 2006:2059). These hours are said to represent approximately 85% of the child's waking experience. By the end of the project, the size of the corpus will have increased by six-fold. It is claimed that this collection of data "constitutes the most comprehensive record of a child's development made to date" (Roy et al. 2006:2059) which must surely be true. While various lab-based experimental methods are available to explore aspects of language acquisition by babies and infants, a corpus-based approach remains a significant part of such research. Conversational analysis has employed similar methods to language acquisition, requiring naturalistic data, and a corpus is the primary data type for this sub-field of linguistics – see, for example, the corpora of spoken interaction at TalkBank (http://talkbank.org).

Psycholinguistics also stands out as one field of linguistics for which a corpus has been relevant, at least for some kinds of psycholinguistic research. I refer here to studies which have explored or rely upon "frequency effects". Frequency effects refer to behavioral patterns which correlate with, or are somehow dependent upon, the degree of familiarity that a form has. An obvious measure of degree of familiarity, though not without its difficulties, is the frequency of occurrence of the form in a corpus. One kind of linguistic behavior where frequency effects can be seen concerns "reaction time" or "response latency", i.e., how long it takes the subject to decide if the word is a real word or a nonsense word. It has been claimed that frequent words take less time to access than less frequent words, though other factors may also play an important role. Indeed this claim has been referred to as the "*sine qua non* of the variables that affect basic word recognition" (Burgess and Livesay 1998:272, cf. also Chiarello 1988). "Frequency effect" could just as well be called "corpus effect" since behind the estimations of frequencies lies a corpus. Even some of the other factors that have been claimed to play a role in reaction time experiments still make use of a corpus. So, for example "contextual diversity" (the number of different texts that a word is found in) is claimed to be a relevant factor in making lexical decisions (Adelmann et al. 2006). Without denying that the exact relation between word frequency in a corpus and subjects' decision times about words is complex (see Forster 2007: 42-44 for further discussion), it is nevertheless clear that frequency of usage, hence a corpus, has always been regarded as a critical factor in such research.

Although the notion of frequency effects is most familiar from psycholinguistics, one could interpret almost all of corpus-based work in linguistics as exploring, to one degree or another, frequencies and, implicitly, frequency effects. This applies, above all, to the research of Joan Bybee, her colleagues, and students. A recent volume, Bybee (2007), contains key publications of hers ranging over almost 30 years and documents a consistent research agenda of exploring the role of high vs. low frequency of usage in many areas of language. These areas include the diffusion of sound change, morphological change, grammaticalization, degrees of constituency in syntax and many others. Bybee (2007:9-18) distinguishes three sub-cases of frequency effects: the conserving effect (repetition reinforces memory representations for linguistic forms and makes them more accessible when retrieving them); the reducing effect (frequency of use leads to more efficient speech articulation which in turn leads to more phonetic assimilation and reduction of the linguistic output); autonomy (the likelihood of a word being stored as a whole and separate unit in the mental lexicon, with frequency of usage being one contributing factor). The word "corpus" does not appear in any of the titles of Bybee's publications in this collection, but it could have, since corpora underlie the estimations of frequencies. The corpora used by Bybeee have varied over the 30 years, beginning with the 1 million word BROWN corpus of English, but extending to a variety of written and spoken corpora of languages other than English.

Two recent articles based on frequency effects (Lieberman et al. 2007 and Pagel et al. 2007) deserve special mention, having appeared in *Nature* in 2007, testifying to a general scientific interest in frequency effects. Lieberman et al. (2007) studied the rate at which a language grows more "regular", with the regular vs. irregular verb paradigms of Old English, Middle English, and Modern English used as a case study. The authors reach the conclusion that irregular verbs undergo regularization, i.e. they conform to the *–ed* type of past tense formation, at a rate that is inversely proportional to the square root of their usage frequency. "Usage frequency" was computed on the basis of the modern reflexes of the words studied, using the CELEX corpus. Pagel et al. (2007) established the rate of replacement of lexical items for 200 meanings and then compared the rates of replacement of the words with their usage frequencies (again based on modern usage). The authors carried out the comparison on data from four

languages: English, Spanish, Russian, and Greek, and found an inverse relationship between frequency and replacement rate: the more frequent a meaning is, the slower its rate of replacement. Furthermore, different parts of speech have different rates of replacement for a given frequency of use. Prepositions and conjunctions evolve most quickly, for example. For this study, the corpora were large corpora from the four languages, and included the British National Corpus in the case of English. A third article (Fitch 2007) in the same issue of *Nature* introduces and comments on the Lieberman et al. and the Pagel et al. articles.

Sociolinguistics has often relied upon collections of texts or spoken material as the basis for claims about variation of linguistic features across different social groups. A classic work of sociolinguistics, Labov's (1972) *Sociolinguistic Patterns*, makes use of a number of different methodologies when it comes to the collection of data, but the collection of samples of connected speech (= a corpus) is certainly an important one and the basis for many kinds of frequency counts. An outstanding modern example of a sociolinguistically oriented corpus is the North Carolina Sociolinguistic Archive and Analysis Project, NC SLAAP (http://ncslaap.lib.ncsu.edu/index.php), described more fully in Kendall (2007). This collection, accessed through the internet, provides an exciting new model for the integration of audio and text from a variety of languages (predominately American dialects in North Carolina and the southeastern United States). At present, it contains over 430 hours of audio, with 26 hours of audio and 250,000 words transcribed at the time of writing. The accompanying tools will allow for searching restricted by speaker and demographic features.

Linguists might wish to distance themselves from the field of language teaching, but if we allowed ourselves to consider this field, then we see here, too, a long-standing interest and much activity in corpus-based research. (cf. McEnery and Wilson 2001:4). Hunston (2002:170-216) reviews some of the history of the field, within the British context, and presents a good discussion of the merits of the use of corpora in language teaching. An early pioneer in this field who deserves to be cited is H. V. George, once a Director of the English Language Institute at Victoria University of Wellington (New Zealand) George carried out (presumably by hand without the aid of computer technology) an impressive corpus-based study of verb patterns in English, published as George (1961). This work reported frequencies of inflectional categories of English verbs in a number of genres (plays, conversation, novels, travel, newspapers, factual). His inflectional categories were exceptionally fine-grained, distinguishing, for example, the 'simple past habitual', 'modal (of modesty)', 'simple past irrealis', and 'simple past neutral' of the *V-ed* form. George's research, which was directed towards improving course design for the teaching of English, stands out as a highly original and unique early contribution to the study of English language usage.

Dictionary-making might also be seen as outside the purview of linguistics by some, but in this area, too, corpora have always had a role to play. This is no less true of the last few decades where dictionaries based on electronic corpora have appeared. The Collins COBUILD English Dictionary (Sinclair 1995) is a prime example of such corpus-based dictionary-making, based as it is on a 200 million word corpus, the Bank of English (now grown to more than twice that size) at the University of Birmingham. A good summary of the corpus-based features of this dictionary can be found at http://www.athel.com/cobuild/cob2.html. Corpus-based dictionaries like the COBUILD dictionary offer a number of advantages, such as placing words in frequency bands, capturing contemporary usage effectively, and using natural examples to illustrate context of usage. The Collins COBUILD English Grammar (Sinclair 1990) is the grammar counterpart to this dictionary.

2.2 Corpora and syntax

From the above, one might form an impression that corpora have played their part in all areas of modern linguistics equally. This would be wrong. As is well known, corpora have been conspicuously lacking in some areas, particularly those that some linguists regard as "core" areas. This applies above all to the study of syntax and semantics in so far as linguists are wont to adopt the standpoint of Chomsky:

(1) Linguistic theory is concerned primarily with the ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors in applying his knowledge of language in actual performance. (Chomsky 1967:3-4):

This dictum of Chomsky's is so well known that it does not require any further interpretation. Suffice to say that, to the extent one feels bound by it, then to the same extent one will eschew the study of actual usage of language. Chomsky's influence on the development of linguistics, especially syntax, has been profound and, of course, not just

because of the dictum quoted above. Without detracting from the profundity of his contribution, it may still be pointed out that not all linguists found this dictum convincing and there have been whole schools of linguistics which did not accept it, such as the Birmingham school, Hallidayan linguistics, etc. To cite one example of resistance to the Chomskyan dictum, the eminent American lexicologist Allen Walker Read had cause to reflect upon it as follows:

(2) You will note how far removed [the dictum quoted in (1)] is from the realities of everyday speech . When do you have an "ideal" speaker-listener? When is any speech-community "completely homogeneous"? What person in the wide world knows even his own language "perfectly"? To escape those deadening restrictions, many of us, as Chomsky's authority became more over-bearing, moved over to the field [sociolinguistics] that allowed us to deal with language as found on the tongues of real people. (Read 1982:17)

Some of the history of Chomskyan linguistics and its stance towards corpora can be found in McEnery and Wilson (2001:5-12) and it is not necessary to repeat that here. I might point out, though, that even among linguists who were working towards developing generative linguistics, one can find corpus-based research. A case in point would be an important contribution by Huddleston (1971): *The Sentence in Written English: A Syntactic Study Based on an Analysis of Scientific Texts*.

Good textbooks are effective at distilling the essence of the subject matter they deal with and I find it useful to turn to two textbooks of modern linguistics to appreciate how data has been traditionally construed in syntax. The first one is Soames and Perlmutter (1979). Its title is already instructive as to what lies at the core of the book: *Syntactic Argumentation and the Structure of English*. The book is, above all else, about *argumentation*, and Soames and Perlmutter have written an outstanding book from that point of view. Argumentation is seen as the real challenge and the path to greater insight and mastery of the scientific method (cf. Soames and Perlmutter 1979:xi). Data, on the other hand, is seen as presenting the lesser challenge. Indeed, the authors point out (p.xi): "An important advantage of linguistics in this respect is that its data is generally much more accessible than data in other sciences and typically can be obtained without time-consuming experiments". And in the first chapter introducing a problem and the process of hypothesis testing, the authors begin with the observation (p. 8): "The discovery of the rules of grammar begins with an examination of linguistic data". Notice that the discovery of the rules of grammar does *not* begin with anything like the collection of data, the verification of data, the role of spoken data vs. written data, or the replicability of the data, all of which would also contribute to the lasting value of a "scientific" approach, notwithstanding the application of the "scientific method". The data somehow just takes care of itself as long as you have a native speaker.[1] The sentences which constitute the data are evaluated, for the most part, in one of two ways: either they are assigned no asterisk, in which case they are presumably perfectly acceptable data (this point is not actually discussed, as far as I can tell), or they each have a single asterisk, in which case they are "in some way deviant" (p. 20). Given the size of the textbook (over 600 pages), it is quite an impressive feat that the authors are able to rely upon this single dichotomy without resource to additional devices, though occasional use is made of ? (p. 294), ?? (p. 294), *? (p. 241) and ** (p. 283), without further explanation. A second textbook, and one which I admire greatly for its pedagogical effectiveness, is Napoli (1996). A reader can be grateful for Napoli's directness, I suppose, when it comes to explaining how the data was collected: "When we look at data from English, we will use the data from a single speaker: me." (p. 296). Or, alternatively: "… we are using my speech since I'm writing this book and my speech is available to me" (p. 296).

2.3 The empirical turn in modern linguistics

The increasing interest in usage-based data is true of the field of linguistics and is by no means restricted to corpus linguistics. I see it as part of an "empirical turn" in contemporary linguistics by which I mean a shift towards a greater interest in many types of data in the quest for a fuller understanding of language. The change must be seen as a "turn" rather than a complete "paradigm shift", but many linguists are no longer content to define their research agenda to conform to the Chomskyan dictum cited above. One of the most succinct, and (I believe) accurate, statements relating to this shift can be found in the abstract (accompanying the internet announcement of the

[1] The relative absence of discussion about the quality, robustness, replicability etc. of introspective data in textbooks like Soames and Perlmutter (1979) is symptomatic of the blindness to the problem of data in the whole field of linguistics. I agree fully with Kepser and Reis (2005:1-2) in recognizing the need for making linguistic evidence a matter of central importance: "…*linguistic evidence* is an extremely important topic as well as a challenging problem for linguists of all persuasions".

publication on the LINGUIST list) of Kepser and Reis' (2005) *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives* volume:

(3) The renaissance of corpus linguistics and promising developments in experimental linguistic techniques in recent years have led to a remarkable revival of interest in issues of the empirical base of linguistic theory in general, and the status of different kinds of linguistic evidence in particular. Consensus is growing (a) that even so-called primary data (from introspection as well as authentic language production) are inherently complex performance data only indirectly reflecting the subject of linguistic theory, (b) that for an appropriate foundation of linguistic theories evidence from different sources such as introspective data, corpus data, data from (psycho-)linguistic experiments, historical and diachronic data, typological data, neurolinguistic data and language learning data are not only welcome but also often necessary.

It is in particular by contrasting evidence from different sources with respect to particular research questions that we may gain a deeper understanding of the status and quality of the individual types of linguistic evidence on the one hand, and of their mutual relationship and respective weight on the other. (LINGUIST list,
available at http://listserv.linguistlist.org/cgi-bin/wa?A2=ind0601b&L=linguist&P=13275)

Presumably the piece is written by the authors, but in any case it very succinctly sums up the authors' own introduction to the volume and describes well the desire on the part of many linguists for more diverse kinds of data in trying to understand the full complexity of language, as well as increasing unease about any simplistic acceptance of one speaker's judgments about the acceptability of example sentences. It is worth noting that the Kepser and Reis book is published in the series *Studies in Generative Grammar*.

The "empirical turn" is not just about an increased respect for, and use of, corpora; it refers to a broader shift involving increased use of many types of data. Nevertheless, increased corpus-based research is a key part of this shift. There are many indications of this shift, including the observation by the current Editor of *Language* concerning the contents of the journal: "…we seem to be witnessing…a shift in the way some linguists find and utilize data – many papers now use corpora as their primary data, and many use internet data" (Joseph 2004:382). A number of new journals have appeared which make corpus evidence their focus, such as the *International Journal of Corpus Linguistics*, *Corpus Linguistics and Linguistic Theory* and *Corpora*. New conferences have emerged where corpus-based research is central, such as ICAME, the American Association of Corpus Linguistics, Quantitative Investigations in Theoretical Linguistics, and many others. Specialized workshops and seminars on corpus-related activities are common, such as the recent conference and master class in *Corpus Methods in Linguistics and Language Pedagogy* at the University of Chicago, held on March 26-30, 2008. A highly acclaimed grammar of English, anchored in the methodology of contemporary theoretical linguistics, Huddleston and Pullum (2002), also acknowledges the role of corpus evidence. The authors describe their data sources as follows:

(4) The evidence we use comes from several sources: our own intuitions as native speakers of the language; the reactions of other native speakers we consult when we are in doubt; data from computer corpora (…). and data presented in dictionaries and other scholarly work on grammar. We alternate between the different sources and cross-check them against each other, since intuitions can be misleading and texts can contain errors. (Huddleston and Pullum 2002:11)

3. Low-level patterns

3.1 Verb inflections

For some years now, my colleague Sally Rice and I have been exploring distributional, collocational, syntactic, and semantic patterns which are associated with specifically inflected forms, as opposed to lemmas (Newman and Rice 2004, 2006a, 2006b; Rice and Newman 2005, 2008). Studying linguistic behavior at the inflected level of words, as opposed to generalizing linguistic behavior at the lemma level, finds support from a number of other areas of linguistic research:

(5) *Grammaticalization studies*. As is well known, grammaticalization can affect particular inflected forms, but not the whole lemma. Consider, for example, the emergence of *going to* as a progressive marker in English, based on a present participial form of the lemma GO.[2] Or consider the emergence of *used to* as a past habitual marker in English, based on the past tense form of the lemma USE.

Reaction time studies. Inflected forms are associated with their own particular reaction times in psycholinguistic experiments where subjects are required to judge whether a form constitutes a word or not. It is not the case that all inflected forms of a lemma are associated with one reaction time. See, for example, Kostić and Havelka (2002) who discuss different reaction times for different person and number forms of Serbian verbs in the future tense (cf. also the discussion of inflectional forms of Serbian noun paradigms in Kostić and Mirković 2002). The issue is not whether there are different reaction times in this literature; the issue is how to account for them.

First language acquisition. Words are acquired in particular inflectional forms and do not suddenly appear in all inflectional categories of a lemma. Morphological marking, e.g., past tense or progressive aspect, attaches to specific lexical items and only much later does it generalize to all stems (cf. Tomasello 1992, 1997:350). An inflected form like *spilled* or *crying* is acquired in its own right before the full lemmas SPILL or CRY are acquired.

Stylistics. Inflectional differences can be significant indicators of genre differences. Greater use of past tense, as opposed to the present tense, for example, is indicative of a narrative style of discourse (cf. Biber 1988 and Biber, Conrad, and Reppen 1988:135-171).

As a way of introducing the reader to this line of study in corpus-based research, let us first consider the data in Figure 1 which summarizes the relative frequencies of the inflected forms of selected English verbs in the 4.2 million word "spoken demographic" (= casual conversation) sub-corpus of the British National Corpus (BNC).[3]

[2] I use small caps to represent the lemma (GO) and italics for inflected forms (*go, goes, gone, went* etc.).

[3] The BNC, i.e., CLAWS tags relevant to this study are: VVB (verb base), VVZ (3SG use of present tense), VVI (infinitive), VVD (past tense), VVG (*-ing* form), VVN (past participle).

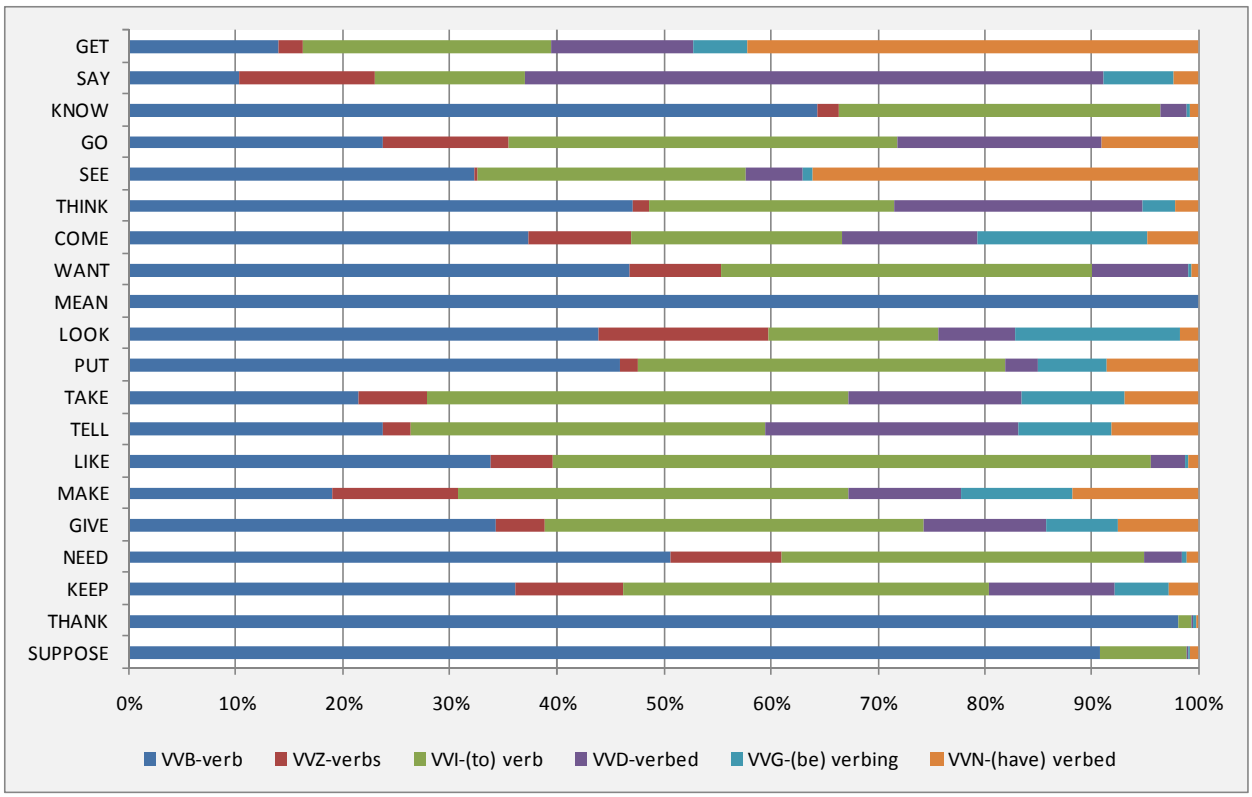

Figure 1. Relative frequencies of the inflected forms of selected verbs in the casual conversation sub-corpus of the BNC (Rice and Newman 2005)

It is immediately obvious from Figure 1, as one might suspect even without the benefit of the quantitative data, that no two verbs behave in an identical way. Nevertheless, no amount of speculation could produce the kind of detail seen in Figure 1. Clearly, there is much fluctuation of the relative proportion of inflected types across verbs. In certain cases, one inflectional category accounts for almost all of the tokens, e.g., the base form of MEAN and THANK (VVB in the BNC tagset). In the case of SUPPOSE, fully 90% of tokens are accounted for by the base form. With these three verbs, one can easily imagine which uses in conversational English might lie behind the facts, in particular, the recurring phrases *I/you mean*, *thank you*, and *I suppose*. Such phrases have taken on particular functions in conversation resulting in easily recognizable, fixed expressions marking features of discourse, features which are not present in the same way or to the same extent in past tense or present perfect uses or with other person/number subjects. GET and SEE, on the other hand, have the highest proportions of past participle forms. The past participle *got* has a well-studied range of unusual uses, including usages such as *I've got a headache,* where the present perfect form is arguably functioning simply as a semantic present tense, equal semantically to *I have a headache*. But why should the past participle *seen* have a similarly high proportion of occurrence? Many of these occurrences of *seen* are found in *I've seen, I have seen, I haven't seen,* and *have you seen*. Unlike the *have/has got* construction, these are less well studied as constructions and, on the surface, completely ordinary uses of SEE, not idiosyncratic or grammaticalized uses. And yet their over-representation in usage requires some explanation and warrants further investigation. There are other intriguing comparisons which emerge from this chart. Comparing SEE and LOOK, for example, one finds with LOOK a relatively high proportion of the present tense form *looks*, whereas the *sees* form of *see* is almost completely lacking. Presumably, the impersonal construction *it looks like/as if* and the absence of *it sees/like as if* contributes to this imbalance.

In some of these cases, we can fairly readily identify the kind of construction which helps us to make sense of an over-representation of a particular inflected form, but this is not always the case (as with the over-representation of the past participle use of SEE). It is precisely in these cases where the corpus reveals patterns which take us beyond what intuition alone can tell us. One particular corpus-linguistic tool, Sketch Engine (Kilgarriff and Tugwell 2001; Kilgarriff, Rychly, Smrz, and Tugwell 2004) offers the user a particularly useful tool for the examination of exceptional behavior of inflected forms of English nouns and verbs. Sketch Engine signals when a particular

inflected form has a "salient" distribution. So, for example, the *–ing* (VVG in the BNC tagset) form *waiting* has a relatively high frequency of occurrence within the set of inflected forms of the verb lemma WAIT. *Waiting* occurs 8,053 times in the BNC, which is 40.62% of all forms of the lemma WAIT which occurs 19,824 times. This puts WAIT in the top 10% of verbs ranked by descending relative frequency of their *–ing* forms (as calculated by Sketch Engine), hence *waiting* is a "salient" inflected form of WAIT. This key result is indicated along with other results from a Word Sketch query without the need for any calculations on the part of the user.

A pair of verbs which have been "stranded" in certain inflections are RID and ALLOW. The graphs in Figure 2a and 2b show the frequencies per million words of these verbs in their various inflections, again in the casual conversation sub-corpus of the BNC. It is virtually only as past participles that we encounter these verbs – something that speakers can have some vague intuitions about, but not to the extent that one could "know" this distribution in this detail. Unlike the verb RUMOUR, which appears only as a past participle in the BNC, both RID and ALLOW are possible in different inflections and speakers' intuitions may not be quite as strong in imagining the usage possibilities of these verbs. There is, in fact, one single instance of a (reflexive) *–ing* form of RID in the casual conversation sub-corpus of BNC, compared with 736 instances of the past participle: *Wolves in fact have done more than most to provide evidence that the game is ridding itself of violence.* It may be possible for some speakers to accept a simple (non-reflexive) use of *ridding*, too, in constructed sentences such as *God is actually ridding the world of all evil, though it may take a while*. Inflected forms other than the past participle may, then, be possible, but the reality of usage is that there is an overwhelming predominance of the past participle.

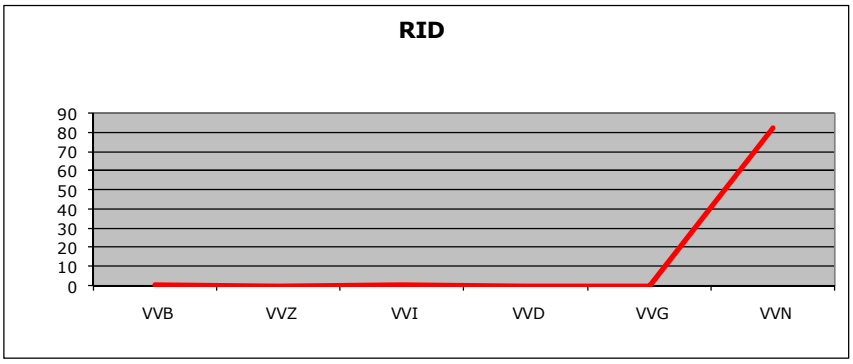

Figure 2a. Frequency per million words of RID in the casual conversation sub-corpus of the BNC (Rice and Newman 2005)

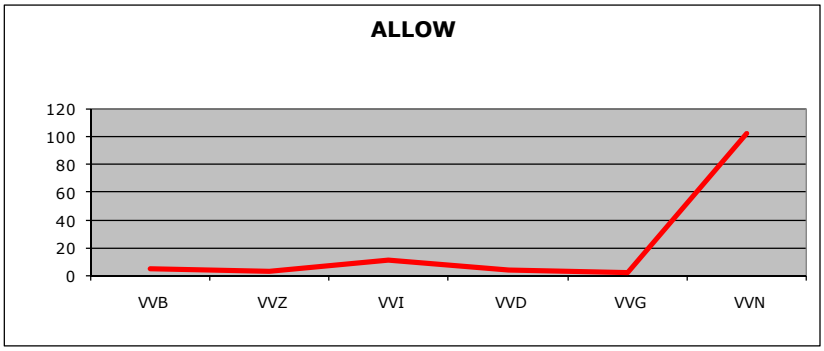

Figure 2b. Frequency per million words of ALLOW in the casual conversation sub-corpus of the BNC (Rice and Newman 2005)

To understand better the usage facts surrounding these verbs, it is necessary to look beyond the individual verbs and to consider the larger constructions of which they form part. In the case of RID, for example, about half of all instances involve some variant of *get rid of*, and in the case of ALLOW, about two-thirds of all instances involve a variant of *be allowed (to have).* This points to the need to extend the discussion to the construction level, rather than restricting ourselves just to the word level. In the case of *get rid of,* there is, in fact, considerable inconsistency in the automatic part-of-speech tagging of the word *rid* in this construction. Table 1 summarizes the results from a few

taggers with respect to sample sentences involving *be rid of* and *get rid of*. This inconsistency is a further indicationof the unusual grammatical status of *rid* which, arguably, could be analyzed as either an inflectionally restricted verb or as an adjective. The corpus results, by themselves, do not solve the question of what the grammatcal status of *rid* should be in these sentences, but the corpus successfully alerts us to a skewing in the inflectional profile of *rid*, warranting further investigation into the potential grammaticalization of this form.

	*I am now completely **rid** of such things.*	*You are well **rid** of him.*	*I got **rid** of the rubbish.*
CLAWS tagger[4]	past participle	past participle	past participle
Infogistics[5]	verb base	verb base	past participle
FreeLing 2.0[6]	adjective	verb base	past participle
Brill-based GOTagger 07[7]	adjective	adjective	adjective

Table 1. Results of automatic part-of-speech tagging of *rid* by four taggers

3.2 Subject and Verb

A natural way to extend the scope of the discussion of inflected forms of verbs is to consider interactions between inflected forms of verbs and the person and number marking of the subjects that occur with them. English verbal inflections indicate this to a limited extent through the contrast between the base form (e.g., *sing*) and the 3rd present tense form (e.g., *sings*). In this section, I proceed to examine interactions between verbs and subjects at this level of inflectional detail. Following Rice and Newman (2005), I illustrate the kind of interactions we find with the verbs THINK, KNOW, MEAN, WANT, and SAY.

In order to show the interaction between subject and verb, I use the convention found in Figure 3. In this set of graphs, the left-hand graph continues the style of Figure 2a-b, showing the relative frequencies per million words of each inflected verb form. On the right-hand side is a graph indicating the relative proportions of the different types of subject occurring with each of the verb inflections (or at least the main ones attested). Numbers on the line graphs indicate the rank of a verb in a specific inflection, by descending frequency, in the set of all verbs with that inflection in the corpus. As can be seen, these verbs are among the most frequently occurring verbs in the corpus, though there can be huge discrepancies between the inflectional categories, as indicated already in Figure 1.

Figure 3 gives us an opportunity to examine, in fine detail, differences between verbs which, on the surface, would seem quite similar, e.g., THINK and KNOW. Both verbs have to do with mental events/ states, but they differ in interesting ways. The past tense *thought* is more frequent than the past tense *knew*, for one thing. Notice, too, that 1SG is dominant as the subject category for all inflections of THINK, whereas this is not the case for KNOW. Combining these observations, we must conclude that *I thought* has a somewhat different status from that of, say, *I knew*. Possibly, the use of *I thought* in utterances like *I've just put the kettle on. Oh! I thought you might be ready for one* (from the corpus) is relevant. In such cases, *I thought* refers to a very recent state of mind, extending into the time of the utterance and is almost equivalent semantically to the explicitly present tense forms *I imagine* or I *assume*.

A comparison of THINK, KNOW, and WANT reveals further patterns which could hardly be intuited. One sees a declining proportion of 1SG subjects as one proceeds through the set THINK - KNOW – WANT, matched by an increasing proportion of 3SG subjects. Thus, there is a kind of inverse-like relationship between 1SG and 3SG in the case of these three verbs. While all three verbs have to do with mental events/states, there are certainly degrees of difference in the extent to which they are predicated of the self versus others, with THINK highly constrained to the self, whereas WANT is most easily predicated of others.

MEAN and SAY present further, quite different profiles in that we find unusual skewing towards inflectional categories – the verb base for mean (the third most frequent verb base in the corpus), and the past tense and the 3SG

[4] http://ucrel.lancs.ac.uk/claws/trial.html

[5] http://www.infogistics.com/posdemo.htm

[6] http://garraf.epsevg.upc.es/freeling/index.php?option=com_content&task=view&id=18&Itemid=47

[7] http://uluru.lang.osaka-u.ac.jp/~k-goto/use_gotagger_e.html

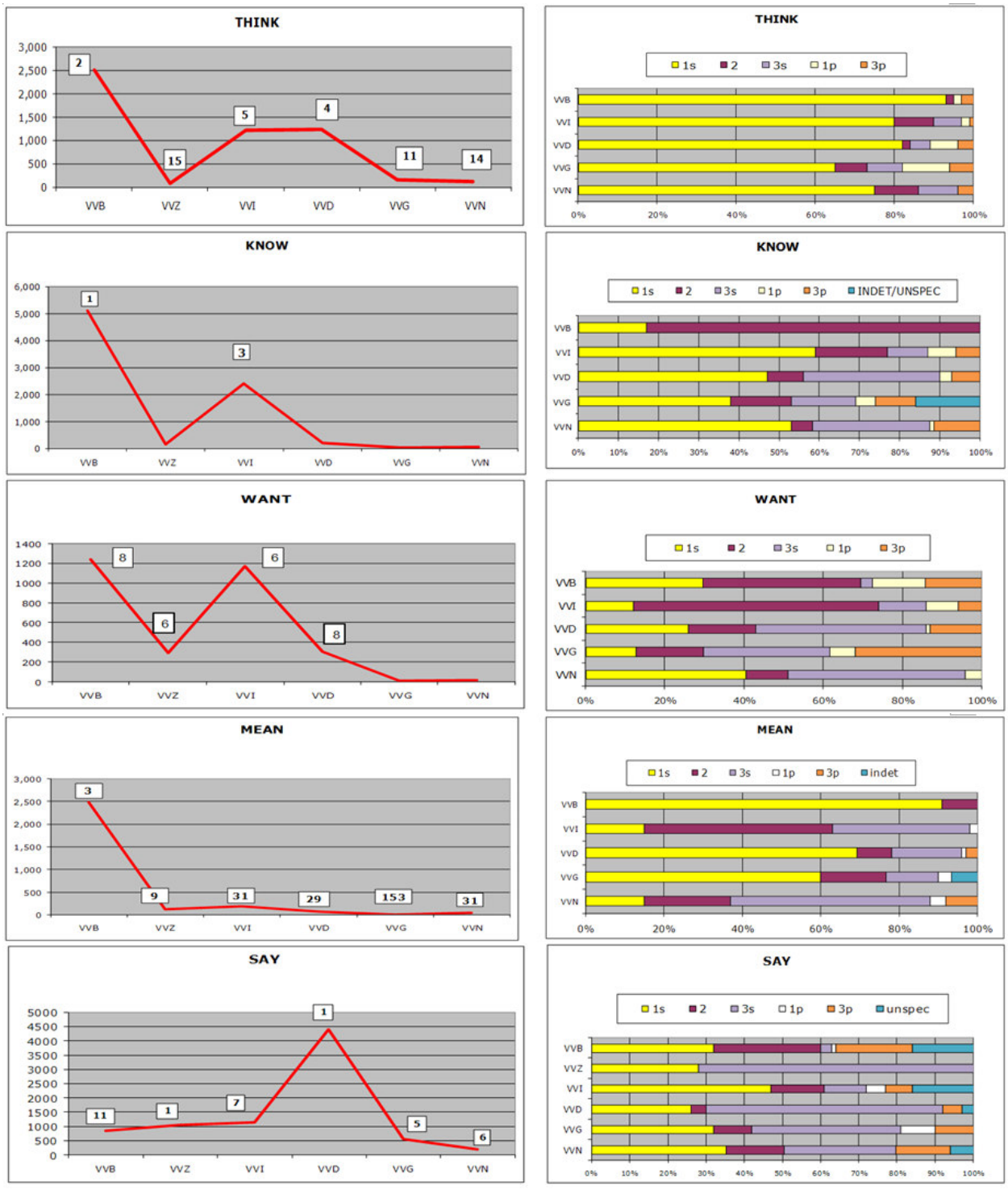

Figure 3. Combined representations of frequencies of inflected verb forms and relative proportions of their subject types for 5 verbs in the casual conversation sub-corpus of the BNC (adapted from Rice and Newman 2005). Numbers on the line graphs indicate the rank of a verb in a specific inflection, by descending frequency, in the set of all verbs with that inflection in the corpus.

present tense for SAY (the most frequent verb in these inflections in the corpus). Superficially, MEAN would seem to be semantically similar to THINK or KNOW, but again there are subtle differences in the relative proportions of subject types. Note, for example, the relatively small proportion of 1SG with the past participle, as in *I've meant, I have meant, I had meant*. One should also note the occurrence of a colloquial *I says* in utterances such as *He says why? I says because things go missing*, accounting for the otherwise odd use of 1SG subjects with 3SG inflectional forms of SAY.

3.3 Verb *and* verb

It is not just the verb and its arguments such as subject, object etc. which can be examined at this level of granularity. All kinds of co-occurrence relationships can be the focus of such studies: contiguous, non-contiguous, syntactic, lexical, lexico-syntactic etc. Here I give just one example of extending this kind of analysis to coordinated verbal units (*V and V*), adapted from Newman and Rice (2004).

"Posture verbs" are an interesting object of linguistic study as evidenced by the range of findings discussed in Newman (2002a). Such verbs prove to be of interest for a variety of reasons, including their polysemy, patterns of grammaticalization, and metaphorical extensions. As a way of approaching a corpus-based study of such verbs in English, we begin with a simple listing of frequencies of posture verbs in a modest corpus of English, the 1 million word (written) BROWN corpus, based on 1960's American English. Conveniently, a large part of the BROWN corpus has been tagged semantically and this is useful for present purposes. Verbs such as SIT, STAND, BEND, etc. can have multiple meanings and so it useful to rely upon semantic tags if we wish to specifically identify "position" or "posture" senses. which are irrelevant to the present discussion and these errant senses must be identified and removed from the database. The most obvious intrusion of unwanted uses concerns the 'tell a lie' sense found with LIE. In addition, the past tense *lay* is identical to some forms of the transitive verb LAY. The documentation accompanying WordNet provides statistics on the frequency of lemmas restricted to a particular meaning in this corpus (Fellbaum 1998). Table 2 lists the frequency of occurrence of a selection of English verbs associated with the rest position, along with a specific WordNet meaning and the number of times that word with that particular meaning occur in the semantically tagged part of the corpus.

SIT 'be sitting'	47
STAND 'be standing, be upright'	43
LIE 'be lying, be prostrate, being a horizontal position'	35
HANG 'be suspended or hanging'	12
LEAN 'incline or bend from a vertical position'	7
SQUAT 'sit on one's heels'	5
KNEEL 'rest one's weight on one's knees'	2
LOUNGE 'sit or recline comfortably'	2
CROUCH 'sit on one's heels'	2
STOOP 'bend one's back forward from the waist on down'	2
BEND 'bend one's back forward from the waist on down'	1
PERCH 'sit, as on a branch'	1
SPRAWL 'sit or lie with one's limbs spread out'	1

Table 2. Frequencies of posture verbs with specific senses in the semantically tagged files of the BROWN corpus

As simple as this exercise is, it successfully reveals to us the special status of SIT, STAND, and LIE within this set of verbs, since they are set apart in terms of their relatively high frequency of occurrence. Interestingly, the counterparts to these three verbs play a peculiarly interesting role in many languages. So, for example, it is just these three verbs which have been extended to grammaticalized, tense/aspect uses in some languages and it is just this set of verbs which from the basis of a 3-way noun classification system in the Amerindian language Euchee (see Newman 2002b).

In this section I will focus on corpus patterns based on SIT, STAND, and LIE. In keeping with the approach adopted in Sections 3.1 and 3.2, I will continue to explore properties of these three verbs at the level of specific inflections. Furthermore, I will limit the discussion to coordinated structures involving variations on *Verb and Verb* coordinated structures. This focus relates to the role that such structures have been shown to play in grammaticalization. So, for

example, in English the coordinated structure *try and (do something)* has a special status in the way meanings of the two verbs (TRY and the second verb) coalesce into one conceptualized event, as opposed to two chronologically or conceptually unrelated events. Here, too, we see the importance of specific inflectional forms of a verb in the grammaticalization process, since this construction is virtually limited to the infinitival form of the verbs. Lichtenberk (2002:274-282) discusses a Manam (Oceanic) construction involving the grammaticalization of a verb *eno* 'lie' which functions as part of a coordinated structure to convey a frequentative or continuative aspect, as shown in (6):

(6) *i-pile-la-be* *i-eno*
3SG.REALIS-speak-LIMITER-and 3SG.REALIS-lie
'He kept talking.' (Manam; Lichtenberk 2002:280)

Indeed, Kuteva (1999:208) singles out the coordinate structure with *and*-type morphemes as a ''locus-for-change'', facilitating the grammaticalization of posture verbs into aspectual markers. Based on an awareness of the role of posture verbs in coordinated structures in such languages, it seems promising to explore similar structures in English. For this study, it was convenient to use the whole BNC (10 million word spoken and 90 million word written), reflecting British English in the 1970s-1990s. I begin with the *–ing* forms, *sitting, standing*, and *lying* and consider the verb collocates which occur in the coordinated structures *sitting and V-ing, standing and V-ing*, and *lying and V-ing*. With these *–ing* forms, we are dealing with "simultaneous" conjunction. The results are listed in Table 3. The cut-off point for inclusion in this listing was set at >1 for *sitting* and *standing*, but ≥1 for *lying* on account of the discrepancies in numbers of collocate types with the three posture verbs.

POSTURE FRAME	COLLOCATE VERB	TOTAL	POSTURE FRAME	COLLOCATE VERB	TOTAL	POSTURE FRAME	COLLOCATE VERB	TOTAL
sitting and	*watching*	8	*standing and*	*looking*	3	*lying and*	*talking*	1
	listening	6		*staring*	3		*waiting*	1
	waiting	5		*talking*	3			
	looking	4		*balancing*	2			
	reading	4		*bending*	2			
	thinking	4						
	talking	3						
	saying	2						
	staring	2						
subtotal collocates (N>1)		38	subtotal collocates (N>1)		13	subtotal collocates (N≥1)		2
other collocates (N=1)		20	other collocates (N=1)		31			
TOTAL IN CORPUS		58	TOTAL IN CORPUS		44	TOTAL IN CORPUS		2

Table 3. Frequency of collocates of posture verbs in the simultaneous conjunction construction in the BNC (Newman and Rice 2004:367)

Note that, although *sitting and V-ing* and *standing and V-ing* occur in roughly the same numbers (58 and 48), *sitting and V-ing* attracts more recurring collocates, indicating a more entrenched use of *sitting* in this use than *standing*. Furthermore, the recurring collocating *V-ing* verbs fall into semantic groups, especially those of cognition (*thinking, reading*) and perception (*watching, listening, looking, staring*). Many verbs are associated with durative states or activities (*watching, waiting, reading, staring,* etc.). With *standing and V-ing*, we find traces of these same meanings, though there is an additional physical dimension in evidence (*balancing, bending*). The presence of these two verbs in the set of recurring collocates with *standing and . . .* is especially interesting in that it mirrors one of the findings of Gibbs, Beitel, Harrington, and Sanders (1994) who found that resistance, center–periphery, balance, verticality, and linkage were the dominant image schemas that subjects identified in their conceptualization of standing. *Lying and V-ing* shows the least number of occurrences as a construction type (just 2), though one can still

discern a durative component of meaning to the two collocating verbs (*talking* and *waiting*). Table 4 summarizes these collocation facts.

Posture verb frame	Properties and associations
sitting and . . . V-ing	has robust collocate inventory collocates with verbs of mental activity, cognition collocates with verbs of visual and auditory perception is most strongly associated with extended duration
standing and . . . V-ing	has fewer recurring collocates than *sitting and . . .* collocates with verbs of balance, physical exercise collocates with verbs of visual perception is associated with extended duration
lying and . . . V-ing	has least robust collocate inventory is marked by absence of recurring patterns is associated with extended duration

Table 4. Attributes associated with the English cardinal posture verbs in the simultaneous conjunction construction (Newman and Rice 2004:370)

The collocational facts in Table 4 mirror, to some extent, the grammaticalization facts associated with the posture verbs cross-linguistically. Heine and Kuteva (2002) document the grammaticalization of 'sit', 'stand', and 'lie' verbs into aspect markers with meanings such as progressive marker, present marker, habitual marker, durative marker, etc. (cf. also example (6) above). In English, of course, these verbs have not grammaticalized into auxiliaries, since they retain full lexical meaning and would be analyzed as main verbs. Nevertheless, the collocational facts of the English verbs reveal similar semantic tendencies to what we see in the grammaticalized cases.

Let us now consider the past tense of these verbs in English and, more particularly, let us consider them in their action meanings 'to enter into a sitting/standing/lying state'. In order to zero in on these meanings in the corpus, searches were carried out based on the strings *sat down and V-ed*, *stood up and V-ed*, and *lay down and V-ed*. Table 5 summarizes the main collocational facts concerning these strings which we may describe as "consecutive" conjunction. Immediately, one sees a striking difference in the relative frequencies of these constructions, compared with the results based on the simultaneous construction shown in Table 3. In Table 5, it is the *stood up and V-ed* construction which occurs most frequently by far and it is this construction which has the greatest number of recurring collocates. Proceeding as for the simultaneous conjunction construction, it is possible to extract semantic tendencies of the collocating *V-ed* forms and these are summarized in Table 6. Note that one does not find the tendency towards prolonged, durative meanings as found in Table 4 above. This fact corresponds to a generalization about grammaticalization of posture verbs, namely that it is the stative meanings of posture verbs which are the source for aspectual markers, rather than the active, motion meanings. One difference between *stood up and…* and *sat down and..* lies in the kinds of actions which are performed consequent to the entry into the state. For *stood up and…,* the entry into the state often serves as a prelude to motion to another place (*walked, moved, came* etc.), which is not the case with *sat down and…* . This observation corresponds to the fact that the motion sense of 'stand' may grammaticalize into a consecutive conjunction marker. Gardiner (1957: 391–392), for example, describes a Middle Egyptian *h* 'stand up, rise' used with a past-tense marker to form an auxiliary with the meaning 'thereupon', indicating that there is a further action. 'Stand' verbs have also developed 'become' meanings in a number of Slavic languages, e.g., Russian *stat'* 'become' (cf. Buck 1949:636–637).

POSTURE FRAME	COLLOCATE VERB	TOTAL	POSTURE FRAME	COLLOCATE VERB	TOTAL	POSTURE FRAME	COLLOCATE VERB	TOTAL
sat down and	*watched*	6	*stood up and*	*walked*	30	*lay down and*	*slept*	3
	had	5		*began*	25		*went to sleep*	3
	thought	5		*went*	24		*looked*	2
	waited	5		*said*	19			
	wept	5		*took*	16			
	leaned	4		*moved*	15			
	opened	4		*stretched*	14			
	picked	4		*looked*	13			
	rested	4		*put*	11			
	worked	4		*came*	10			
	helped	3		*shook*	9			
	looked	3		*made*	7			
	sipped	3		*shouted*	7			
	started	3		*smiled*	7			
	tried	3		*gazed*	6			
				held	6			
				faced	5			
				left	5			
				turned	5			
				waved	5			
				done	4			
				kissed	4			
				prepared	4			
				pulled	4			
				reached	4			
				wandered	4			
				crossed	3			
				drew	3			
				opened	3			
				patted	3			
				shuffled	3			
				strode	3			
				yelled	3			
subtotal collocates (N>2)		61	subtotal collocates (N>2)		284	subtotal collocates (N>1)		8
other collocates (N≤2)		200	other collocates (N≤2)		193	other collocates (N=1)		12
TOTAL IN CORPUS		261	TOTAL IN CORPUS		477	TOTAL IN CORPUS		20

Table 5. Frequency of collocates of posture verbs in consecutive conjunction construction in the BNC (Newman and Rice 2004:376)

CPV COLLOCATION	PROPERTIES AND ASSOCIATIONS
sat down and...	presents smaller inventory of collocates than STAND is associated with STATE or ACTIVITY IN A PLACE prefers HUMAN SUBJECTS
stood up...	presents most robust collocate inventory is associated with STATE or ACTIVITY in a place has INCHOATIVE overtones often serves as prelude to MOTION to another place prefers HUMAN SUBJECTS
lay down and…	presents smallest inventory of collocates is associated mainly with relatively PASSIVE STATES OR ACTIVITIES prefers HUMAN SUBJECTS

Table 6. Attributes associated with the English cardinal posture verbs in consecutive conjunction construction

4. Data and theory

In illustrating aspects of corpus-based research, I have focused on the observation of data and the extraction of patterns, at a relatively low level as far as linguistic generalizations are concerned. In the case of the data discussed above, this has involved words at the level of inflected form and patterns that they enter into with other words. Obviously, a corpus-based approach, even limited to the type of research illustrated here, can generate great quantities of such data, not just for English, but for many languages, since comparable data and tools to access the data are now so widely available. Not only is there potentially a vast amount of such data available, but to the extent that the corpora are available to other researchers, so too the research carried out on such data is replicable – an appealing attribute for a scholarly methodology.

It should be equally clear, however, that corpus data in linguistics are not without their problems and that a reliance on corpus data adds to the number of issues we have to deal with when it comes to evaluating data. There is, for example, the issue of how data for a corpus has been collected and understanding what the data represents. Does it represent a recognized genre, a combination of genres, the language of a speech community, etc.? The BNC consists of 90% originally written material and 10% originally spoken material, but one is entitled to question whether those proportions are appropriate for drawing conclusions about (British) "English"? There is the further issue of how we deal with the fact that the corpus presents us with form only, not with any direct indication of what mental processes the speaker and hearer in the speech act context (or author and audience in the case of the written material) actually experienced, what meanings were intended, what meanings were taken from the form. One might achieve replicability of the methods used in relying upon corpus data, but if the corpus itself is unsound in some way then the final result remains unsatisfactory. And where one relies on tags, e.g., part-of-speech tags, to refine searches for data in corpora, there is always a quesiton about the reliability of the tagging process. A corpus-based approach, in other words, has resulted in more *problematizing* of data. However, as I suggested in Section 2, this is a necessary and welcome development in a discipline which has been insufficiently troubled by such issues.

At least two new kinds of problems arise in connection with corpus data: (a) the visualization of data, and (b) the statistical analysis of data. Both of these issues relate to the quantitative nature of corpus-based research. The visualization of data refers to various ways to present many (sometimes thousands) of observations and the visualization problem is to develop the most effective ways to present data in order to discern patterns. Visualizing data was not an issue in traditional, example-bound linguistic analysis where there was, more or less, a single way of presenting data, i.e., a single example (maybe with interlinear glossing) or a small number of such examples to show a contrast or a similarity between the examples. But the results of corpus-based investigations can be viewed and examined in many different ways. Concordance lines showing key words in context, for example, remain a common and favorite way of obtaining a sense of actual context of usage no matter how quantitative the overall approach to the data is. Collocates can be shown in various formats, e.g., tables as used in Section 3.3. A graph-based format, the

"Visual Collator", has been developed by TAPoR which offers quite a different view of collocates.[8] Frequency facts can be viewed in many different formats – as seen in the alternative styles used in Section 3.1 and 3.2 above. The second kind of problem relates to the statistical analysis of corpus-based results. There is no question that statistical techniques are needed, at some point, to help interpret the data and to arrive at results which might satisfy statisticians, i.e., results which can be shown to be "statistically significant". The use of statistically sophisticated techniques has not been very evident in mainstream syntax and semantics in the past, but they are becoming more common as the field becomes more empirically oriented. Many statistical techniques are being employed in corpus-based research, but the corpus presents a relatively new kind of subject matter for the application of statistics and it will take time for "best practice" to establish itself. Even those linguists who are statistically sophisticated seem to have more questions than answers when it comes to deciding what "best practice" is supposed to be. A case in point is Kilgarriff (2005), published in the first volume of the journal *Corpus Linguistics and Linguistic Theory*, and the commentary on it by an editor of the journal (Gries 2005). The two authors address the issue of "null hypothesis testing" in corpus linguistics, and Gries takes up the challenge posed by Kilgarriff's claim that null-hypothesis significance testing (which assumes "randomness") leads to unhelpful or misleading results. Gries (2005:281) suggests ways to deal with some of the problems raised by Kilgarriff, but concludes with the words: "On the basis of these results, it is very difficult to decide what the right quantitative approach in such word-frequency studies may be. What is not so difficult is to notice that much more exploration in these areas is necessary and that the results show how much we may benefit from taking up methodological proposals from other disciplines to refine, and add to, our methods…".

In my illustration of corpus-based research in Section 3, I focused on the potential for new kinds of low-level observations now that we have such widely available corpus data and corpus tools, without elaborating on the way in which such observations inform linguistic theory. Clearly, there is always a danger with such an abundance of data that the data will overwhelm the researcher and that the vast amounts of data will impede rather than promote genuine insight. I see two important ways in which the corpus-based research reported on above can be integrated into a larger linguistic enterprise. One larger enterprise of which the research above forms part is that of constructing a fully worked-out usage-based grammar of English. An outstanding example of a usage-based grammar is the *Longman Grammar of Spoken and Written English* (*LGSWE*, Biber at al. 1999), a grammar drawing heavily upon corpus-based observations. Most of the key points made in this grammar are presented together with their relative frequencies in four genres (conversation, fiction, news, and academic writing), allowing the user to readily appreciate the extent of usage of a word or construction type. The patterns observed in Section 3 are presented in the same kind of spirit, i.e., contributing towards a descriptively more complete account of English language usage.

As laudable as the goals of *LGSWE* are, they remain descriptive in nature, relying upon no more than a traditional grammar understanding of the structure of English. As linguists, we need to strive, ultimately, for a greater understanding of language than simply accepting what traditional grammar has to offer and corpus-based research can contribute to that greater understanding in important ways. In my view, the cause of linguistic theorizing is best served by pursuing, concurrently, a diversity of methodological approaches to understanding language data. A strongly theory-driven approach of the type we are accustomed to in generative linguistics, following the Chomskyan dictum cited above, would be one methodological approach, but it would not be accorded any special status. Rather, it is to the more empirically based approaches to language data that we should now turn, with greater urgency and with greater effort than has been the case. For one thing, we need to correct the over-emphasis on top-down theorizing which has characterized so much of the field of linguistics, to make way for a greater presence of data-driven research. Just as importantly, we must find ways to reconcile the insights into data which emerge from the many data-based approaches we now see in linguistics (corpus data, reaction times to recognizing words, eye-tracking experiments, etc.). Instead of pursuing any one of these approaches in isolation, independent of any awareness about how such data are viewed in alternative approaches, it would seem highly desirable to be constantly

[8] TAPoR is a pan Canadian, multi-institutional assembly of computing infrastructure and expertise, supported by a Canada Foundation for Innovation grant. The University of Alberta TAPoR node is http://tapor.ualberta.ca/. The Visual Collocator is described as follows: "[it] displays collocates of words using a graph layout. Words which share similar collocates will be drawn together in the graph, producing new insight into the text. Any word can be double-clicked to fetch its collocates. Any word can be removed from the graph, and new words can be added using the text field. Additionally, words can be made 'sticky', then dragged around to new positions, creating a user defined layout." (from http://tada.mcmaster.ca/Main/TAPoRwareVisualCollocator)

seeking to reconcile results between the different approaches. Gries, Hampe, and Schönefeld (2005, to appear) are good examples of how we might proceed to explore, simultaneously, corpus and psycholinguistic data and in this way construct solid underpinnings to a larger theory of language.

References

Biber, Douglas (1988). *Variation across Speech and Writing*. New York: Cambridge University Press.

Biber, Douglas, Susan Conrad, and Randi Reppen (1988). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan (1999). *Longman Grammar of Spoken and Written English.* Harlow: Longman.

Buck, Carl Darling (1949) *A Dictionary of Selected Synonyms in the Principal Indo-European Languages*. Chicago and London: University of Chicago Press.

Burgess, Curt and Ken Livesay (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kucera and Francis. *Behavior Research Methods, Instruments, & Computers* 30:272-277.

Bybee, Joan (2007). *Frequency of Use and the Organization of Language.* Oxford: Oxford University Press.

Chiarello, Christine (1988). Lateralization of lexical process in the brain: A review of visual half-field research. In Harry A.Whitaker (ed.), *Contemporary Reviews in Neuropsychology*, pp. 36-76. Berlin: Springer Verlag.

Fellbaum, Christiane (ed.) 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fitch, W. Tecumseh (2007). Linguistics: An invisible hand. *Nature* 449: 665-667 (11 October).

Forster, Kenneth I. (2007). Visual word recognition: Problems and issues. In Gonia Jarema and Gary Libben (eds.), *The Mental Lexicon: Core Perspective*s, pp. 31-53. Amsterdam: Elsevier.

Gardiner, Alan H. (1957). *Egyptian Grammar: Being an Introduction to the Study of Hieroglyphs*. 3rd ed. Oxford: Oxford University Press.

George, H. V. (1961). Report on a verb form frequency count. *Monograph of the Central Institute of English* (Hyderabad) 1.

Gibbs, Raymond W., Jr., Dinara A. Beitel, Michael Harrington, and Paul E. Sanders (1994). Taking a stand on the meanings of Stand: Bodily experience as motivation for polysemy. *Journal of Semantics* 11:231–251.

Gries, Stefan Th. (2005). Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory* 1.2:277-294.

Gries, Stefan Th., Beate Hampe, and Doris Schönefeld (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16.4:635-676.

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

Heine, Bernd and Tania Kuteva (2002). *World Lexicon of Grammaticalization.* Cambridge: Cambridge University Press.

Huddleston, Rodney D. (1971). *The Sentence in Written English: A Syntactic Study Based on an Analysis of Scientific Texts*. Cambridge: Cambridge University Press.

Joseph, Brian (2004). The Editor's Department: On change in Language and change in language. *Language* 80.3:381-383.

Kendall, Tyler (2007). Enhancing sociolinguistic data collections: The North Carolina Sociolinguistic Archive and Analysis Project. *University of Pennsylvania Working Papers in Linguistics* 13.2:15-26. Available at http://ncslaap.lib.ncsu.edu/papers.php#kendall2007a.

Kepser, Stephan and Marga Reis (2005). *Linguistic Evidence Empirical, Theoretical and Computational Perspective*. Berlin: Mouton de Gruyter.

Kilgarriff, Adam (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1.2:263-276.

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

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

Kostić, Aleksandar, and Jelena Havelka (2002). Processing of verb tense. *Psihologija* 35.3-4:299-316.

Kostić, Aleksandar, and Jelena Mirković (2002). Processing of inflected nouns and levels of cognitive sensitivity. *Psihologija* 35.3-4:287-297.

Kuteva, Tanya A. (1999). On ''sit''/''stand''/''lie'' auxiliation. *Linguistics* 37:191–213.

Labov, William (1972). *Sociolinguistic Patterns.* Philadelphia: University of Pennsylvania Press.

Lichtenberk, Frantisek. (2002). Posture verbs in Oceanic. In John Newman (ed.), *The Linguistics of Sitting, Standing, and Lying,* pp. 269-314. Amsterdam and Philadelphia: John Benjamins.

Lieberman, Erez, Jean-Baptiste Michel1, Joe Jackson, Tina Tang, and Martin A. Nowak (2007). Quantifying the evolutionary dynamics of language. *Nature* 449:713–716 (11 October).

McEnery, Tony and Andrew Wilson (2001). *Corpus Linguistics.* 2nd edition. Edinburgh: Edinburgh University Press.

Napoli, Donna Jo (1996). *Linguistics*. Oxford: Oxford University Press.

Newman, John (ed.) (2002a). *The Linguistics of Sitting, Standing, and Lying.* Amsterdam and Philadelphia: John Benjamins.

Newman, John (2002b). A cross-linguistic overview of the posture verbs 'sit', 'stand', and 'lie'. In John Newman (ed.), *The Linguistics of Sitting, Standing, and Lying,* pp. 1-24. Amsterdam and Philadelphia: John Benjamins.

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

Newman, John and Sally Rice (2006a). Transitivity schemas of English EAT and DRINK in the BNC. In Stefan Th. Gries and Anatol Stefanowitsch (eds.), *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis,* pp. 225-260. Berlin and New York: Mouton de Gruyter.

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

Pagel, Mark, Quentin D. Atkinson, and Andrew Meade (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449:713-716 (11 October).

Read, Allen Walker (1982). The contribution of linguistics to the peace-keeping process. *ETC: A Review of General Semantics* 39.1:16-21. Available at http://learn-gs.org/library/awr/39-1-read.pdf.

Rice, Sally and John Newman (2005). Inflectional Islands. Presentation at *9th International Cognitive Linguistics Conference*, Yonsei University, Seoul, Korea. Available at http://www.ualberta.ca/~johnnewm/.

Rice, Sally and John Newman (2008). Beyond the lemma: Inflection-specific constructions in English. Presentation to *American Association for Corpus Linguistics 2008*, Provo, Utah. Available at http://corpus.byu.edu/aacl2008/schedule.asp (paper 129).

Roy, Deb, Rupal Patel, Philip DeCamp, Rony Kubat, Michael Fleischman, Brandon Roy, Nikolaos Mavridis, Stefanie Tellex, Alexia Salata, Jethran Guinness, Michael Levit, and Peter Gorniak (2006). The Human Speechome Project. *Proceedings of the 28th Annual Cognitive Science Conference*, pp. 2059-2064. Available at http://www.cogsci.rpi.edu/csjarchive/Proceedings/2006/docs/p2059.pdf.

Sinclair, John (1990). *Collins COBUILD English Grammar*. London: HarperCollins.

Sinclair, John (1995). *Collins COBUILD English Dictionary*. 2nd, revised edition. London: HarperCollins.

Soames, Scott and David M. Perlmutter (1979). *Syntactic Argumentation and the Structure of English*. Berkeley: University of California Press.

Tomasello, Michael (1992). *First Verbs: A Case Study of Early Grammatical Development.* New York: Cambridge University Press.

Tomasello, Michael (1997). One child's early talk about possession. In John Newman (ed.), *The Linguistics of Giving*, pp. 349-373. Amsterdam: John Benjamins.

John Newman
Department of Linguistics
4-32 Assiniboia Hall
University of Alberta
Edmonton, Alberta
T6J 6H2 CANADA

john.newman@ualberta.ca