CORPUS LINGUISTICS AND THE STUDY OF ENGLISH GRAMMAR

Douglas Biber

Northern Arizona University, USA

Abstract

This paper describes how corpus-based analyses can be employed for the study of English grammar, with a focus on case studies taken from the Longman Grammar of Spoken and Written English (LGSWE). Two major themes are developed: 1) the kinds of unexpected findings about language use that result from corpus-based investigations, and 2) the importance of register for any descriptive account of linguistic variation. Three case studies are presented: one focusing on the use of words (i.e., the most common verbs in English); the second focusing on the use and distribution of grammatical forms (i.e., the relative frequency of simple, progressive, and perfect aspect in English); and the third describing how lexis and grammatical structure can interact in complex ways (i.e., showing how verbs with the same valency patterns can have strikingly different preferences for particular valencies). In all three cases, the paper argues for the centrality of a register perspective, showing how the patterns of use vary dramatically from one register to another.

Keywords: corpus-based analyses, register, linguistic variation, valency patterns

INTRODUCTION: CORPUS-BASED INVESTIGATIONS OF GRAMMAR AND USE

There have been numerous studies of grammar and use over the last two decades, as researchers have come to realize that the description of grammatical function is as important as structural analysis. In most cases, these studies focus on grammatical features that have two or more structural or semantic variants. By studying these features in naturally occurring discourse, researchers have been able to identify systematic differences in the functional use of each variant.

Research of this type became popular in the late 1970's and 1980's. For example, Prince (1978) compared the discourse functions of WH-clefts and *it*-clefts; Thompson investigated word order variation with detached

participial clauses (1983), and adverbial purpose clauses (1985); Schiffrin studied the discourse factors influencing grammatical variation in verb tense (1981), causal sequences (1985a), and discourse markers (1985b, 1987). Other more recent studies of this type include Thompson and Mulac (1991a,b) on the discourse conditions associated with the omission of the complementizer *that*; Fox and Thompson (1990) on relative clauses; and Myhill (1995, 1997) on the discourse functions of modal verbs.

At one level, these studies might be regarded as early corpus-based investigations: they are all empirical studies based on analysis of grammatical features in actual texts. In addition, most of these studies have used both quantitative and qualitative analysis. That is, quantitative techniques are used to determine the distribution of grammatical variants across contexts, while detailed analyses of text extracts are used to interpret the distributional patterns in functional terms.

However, there has often been relatively little concern with the generalizability of the texts used for such analyses. Many of these studies have used a 'convenience' sample: a collection of texts that was readily available to the researcher. The implicit assumption underlying this methodological decision seems to have been that any body of naturally-occurring discourse will illustrate the same patterns of use. However, these text samples have often been small, and more importantly for the present purposes, there has often been no systematic control for register. Some studies are based on a single register; others are based on discourse examples with disregard to register; while only a few incorporate a comparison of use across registers.

More recently, researchers on discourse and grammar have begun to use the tools and techniques available from corpus linguistics, with its greater emphasis on the representativeness of the database, and its computational tools for investigating distributional patterns in large text collections (see Biber, Conrad, and Reppen, 1998 for an introduction to this analytical approach). There have been numerous research papers using corpus-based techniques to study English grammar and discourse. The edited volumes by Aarts and Meyer (1995), Aijmer and Altenberg (1991), and Johansson and Stenström (1991) provide good introductions to work of this type. There are also a number of book-length treatments reporting corpus-based investigations of grammar and discourse: for example, Tottie (1991) on negation, Collins (1991) on clefts, Granger (1983) on passives, Mair (1990) on infinitival complement clauses, Meyer (1992) on apposition, and several books on nominal structures (e.g., de Haan, 1989; Geisler, 1995; Johansson, 1995; Varantola, 1984).

In most cases, corpora are designed to represent some register differences, and thus many grammatical studies based on corpora have a

register component. For example, Tottie (1991) and Geisler (1995) report differences for speech versus writing; Johansson (1995) distinguishes among Press, Fiction, and Academic prose; and Granger (1983) distinguishes among several different spoken registers (including conversation, oration, commentary, interviews). At the same time, other corpus-based studies disregard register distinctions in their studies of grammar and discourse, focusing exclusively on a detailed analysis of contextual factors (e.g., Mair, 1990; de Haan, 1989; Sinclair, 1991).

In the present paper, I take a strong position on the importance of register for studies of grammar and use, arguing that most functional descriptions of a grammatical feature will *not* be valid for the language as a whole. Rather, characteristics of the textual environment interact with register differences, so that strong patterns of use in one register often represent only weak patterns in other registers. Thus, a complete functional analysis must consider the patterns of use across registers.

In the following sections, I illustrate the interaction of grammar, use, and register with corpus-based analyses adapted from the *Longman Grammar of Spoken and Written English* (Biber, *et al.*, 1999). Three cases studies are presented, all focusing on the use of verbs: one dealing with lexical patterns (i.e., the most common verbs in English, Section 2); the second focusing on the use and distribution of grammatical forms (i.e., the relative frequency of simple, progressive, and perfect aspect in English, Section 3); and the third describing how lexis and grammatical structure interact in complex ways (i.e., showing how verbs with the same valency potentials can have strikingly different preferences for particular valencies, Section 4).

The analyses are based on texts from four registers: conversation, fiction, newspaper language, and academic prose. Although these are general registers, they differ in important ways from one another (e.g., with respect to mode, interactiveness, production circumstances, purpose, and target audience). The analyses were carried out on the Longman Spoken and Written English (LSWE) Corpus, which contains c. 40 million words of text overall, with c. 4-5 million words from each of these four registers (see Table 1). All frequency counts reported below have been normalized to a common basis (a count per 1 million words of text), so that they are directly comparable across registers.

Table 1: Composition of the sub-corpus used in the analyses (taken from LSWE Corpus)

	Number of texts	Number of words
Conversation (BrE)	3,436	3,929,500
Fiction (AmE and BrE)	139	4,980,000
News (BrE)	20,395	5,432,800
Academic prose (AmE and BrE)	408	5,331,800

THE MOST COMMON LEXICAL VERBS ACROSS REGISTERS

There are literally dozens of common lexical verbs in English. For example, nearly 400 different verb forms occur over 20 times per million words in the LSWE Corpus (see Biber *et al.*, 1999:370-371). These include many everyday verbs such as *pull, throw, choose,* and *fall*.

Given this large inventory of relatively common verbs, it might be easy to assume that that no individual verbs stand out as being particularly frequent. However, this is not at all the case: there are only 63 lexical verbs that occur more than 500 times per million words in a register, and only 12 verbs occur more than 1,000 times per million words in the LSWE Corpus (Biber *et al.*, 1999:367-378). These 12 most common verbs are: *say, get, go, know, think, see, make, come, take, want, give,* and *mean.*

To give an indication of the importance of these 12 verbs, Figure 1 plots their combined frequency compared to the overall frequency of all other verbs. Taken as a group, these 12 verbs are especially important in conversation, where they account for almost 45% of the occurrences of all lexical verbs. Obviously, any conversational primer that did not include extensive practice of these words would be shortchanging students.

It further turns out that there are large frequency differences among these 12 verbs, overall and in their register distributions. For example, Figures 2 and 3 plot the frequency of each verb in conversation and in newspaper language (cf Biber *et al.*, 1999:374-376). The verb *say* is listed first in these figures because it is common in both spoken and written registers and thus has the highest frequency overall. This is not surprising, given the ubiquitous need to report the speech of others; it turns out that both speakers and writers rely heavily on the single verb *say* for this purpose, usually in the past tense expressing either a direct or indirect quote. For example:

> You said you didn't have it. (conversation, henceforth CONV)

> He said this campaign raised 'doubts about the authenticity of free choice'. (news)

The extremely high frequency of the verb *get* in conversation is more surprising. This verb goes largely unnoticed, yet in conversation it is by far the single most common lexical verb in any one register. The main reason that *get* is so common is that it is extremely versatile, being used with a wide range of meanings. For example:

> Obtaining something:
> See if they can *get* some of that beer. (CONV)
> Possession:
> They've *got* a big house. (CONV)
> Moving to or away from something:
> *Get* in the car. (CONV)
> Causing something to move or happen:
> It *gets* people talking again, right? (CONV)
> Understanding something:
> Do you *get* it? (CONV)
> Changing to a new state:
> So I'm *getting* that way now. (CONV)

Several other verbs are also extremely common in conversation: *go, know*, and to a lesser extent, *think*, *see, come, want*, and *mean*. News, on the other hand, shows a quite different pattern, with only the verb *say* being extremely frequent. However, it should be noted that all 12 of these verbs are notably common in both registers in comparison to most verbs in English. For example, as noted above, verbs like *pull*, *throw, choose*, and *fall* occur only about 50-100 times per million words. Countless other verbs have even lower frequencies. In contrast, the majority of the 12 most common verbs occur over 1,000 times per million words in both conversation and news.

Thus there is a cline in the use of verbs: a few verbs occur with extremely high frequencies; several verbs occur with moderately high frequencies; while most verbs occur with relatively low frequencies. In addition, different registers show strikingly different preferences for particular verbs. For example, the verbs *get, go, know,* and *think are* much more frequent in conversation than in news (see Figures 2 and 3). In contrast, verbs like *add, spend, claim,* and *continue* are much more common in news than in conversation.

SIMPLE, PROGRESSIVE, AND PERFECT ASPECT ACROSS REGISTERS

One of the most widely held intuitions about language use among English-language professionals is the belief that progressive aspect is the unmarked choice in conversation. This belief is sometimes reflected in the overly frequent use of progressive verbs in made-up dialogs (like those found in ESL/EFL coursebooks teaching conversation skills). For example,

> *Conversation from "As I was Saying: Conversation Tactics"*
>
> Doctor : Hello, Mrs. Thomas. What can I do for you?
> Patient : Well, I've *been having* bad stomach pains lately, doctor.
> Doctor : Oh I'm sorry to hear that. How long have you *been having* them?
> Patient : Just in the last few weeks. I get a very sharp pain about an hour after I've eaten.
>
> […]
>
> Doctor : Well, I don't think it's anything serious. Maybe you eat too quickly. You don't give yourself time to digest your food.
> Patient : My husband *is always telling* me that.

As Figure 4 shows, the generalization that progressive aspect is more common in conversation than in other registers is correct. The contrast with academic prose is especially noteworthy: progressive aspect is rare in academic prose but common in conversation. However, the overall register distribution is surprising in that progressive verb phrases are nearly as common in fiction as in conversation, and they are relatively common in news as well.

More surprisingly, as Figure 5 shows, it is not at all correct to conclude that progressive aspect is the unmarked choice in conversation. Rather, simple aspect is clearly the unmarked choice. In fact, simple aspect verb phrases are more than 20 times as common as progressives in conversation. The following excerpt illustrates the normal reliance on simple aspect in natural conversation:

> B: -- What *do* you *do* at Dudley Allen then?
> A: What the school?
> B: Yeah. *Do* you --
> A: No, I*'m*, I*'m* only on the PTA.
> B: You*'re* just on the PTA?

A: That*'s* it.
B: You *do*n't actually *work*?
A: I *work* at the erm -
B: I *know* you *work* at Crown Hills, *do*n't you?
A: Yeah.

In contrast, progressive aspect is used for special effects, usually focusing on the fact that an event is in progress or about to take place. For example:

What*'s* she doing? (CONV)
But she*'s coming* back tomorrow. (CONV)

With non-dynamic verbs, the progressive can refer to a temporary state that exists over a period of time, as in:

I was looking at that one just now. (CONV)
You should be wondering why. (CONV)
We were waiting for the train. (CONV)

A few lexical verbs actually occur most of the time with the progressive aspect in conversation. These include: *bleeding, chasing, shopping, starving, joking, kidding,* and *moaning*. However, the norm – even in conversation – is to express verbs with the simple aspect. In marked contrast to the expectations created by some popular conversational materials, verb phrases such as *I've been having* and *is always telling* are exceptional rather than the rule.

LEXICO-GRAMMATICAL AND REGISTER FACTORS INFLUENCING THE USE OF VALENCY PATTERNS

The above sections have illustrated the unexpected lexical and grammatical patterns of use that can be uncovered by corpus-based research. It further turns out that there are often complex interactions between word sets and grammatical variation. Such **lexico-grammatical** associations usually operate well below the level of conscious awareness, yet they are highly systematic and important patterns of use. In the present section, I illustrate these associations through a comparison of the valency patterns for *stand* and *begin* (see also Biber *et al.*, 1999:380-392; Biber, Conrad, and Reppen, 1998:95-100).

Many verbs take only a single valency pattern. For example, *wait, happen,* and *exist* occur only as intransitive verbs, while verbs like *bring, carry, suggest*, and *find* occur only as transitive verbs. However, there are many other verbs that can occur with multiple valency patterns, such as *eat, try, watch, help*, and *change*.

Stand and *begin* are two verbs that have exactly the same potential for occurring with multiple valency patterns -- both verbs can occur with four different patterns:

Simple intransitive (SV):

For a while he *stood* and watched. (Fict)
A number of adults and children have left the compound since the siege *began*. (News)

Intransitive with an optional adverbial (SV+A):

I just *stood* there. (Conv)
This effort *began* in January of 1981. (Acad*)

Transitive with a noun phrase as direct object (SVO (NP)):

My mom couldn't *stand* it in the end. (Conv)
Mr Hawke's government has *begun* its controversial plan to compensate the three main domestic airlines. (News*)

Transitive with a complement clause as direct object SVO (Comp-cls):

Carrie *stood* shivering in the cold hall. (Fict*)
He *began* to scratch slowly in the armpit of his alpaca jacket. (Fict*)

A traditional grammatical description would simply note that these two verbs occur with the same four valency patterns. However, corpus-based analysis opens up the possibility of a use perspective on such points of grammar. Sections 2 and 3 have shown how the use of words and grammatical features is conditioned by register; the present section shows how the use of grammatical patterns is conditioned by individual words (which is in turn conditioned by register).

In fact, it turns out that the two verbs *stand* and *begin* have strikingly different preferred valency patterns, despite their identical valency potentials. Table 2 shows the proportional use of each verb with each pattern.

Table 2: Proportional use of *stand* and *begin* with intransitive and transitive valency patterns

* Pattern is attested but occurs less than 10% of the time			
** Pattern occurs 10-25% of the time			
*** Pattern occurs 25-50% of the time			
**** Pattern occurs 50-75% of the time			
***** Pattern occurs over 75% of the time			
- Pattern is not attested			
	SV	SV+A	SVO (NP)
SVO (Comp-cls)			
stand			
Conv	***	****	**
-			
Fict	***	****	*
**			
News	***	****	**
*			
Acad	***	****	*
*			
begin			
Conv	**	*	*
*****			
Fict	**	*	*
****			
News	***	***	**
***			
Acad	**	***	**
***			

(Based on Biber, *et al.,* 1999.385; Table 5.5)

As Table 2 shows, these two verbs typically occur with very different valency patterns: *stand* usually occurs as an intransitive verb, often with an optional adverbial, while *begin* is more common occurring with a following complement clause. Further, there are important register differences; for example, the pattern *begin* + complement clause is especially characteristic of conversation, while intransitive *begin* is more likely to occur in news and academic prose.

The predominant use of *stand* as an intransitive verb corresponds to its typical meaning marking a physical state, as in:

> I just sort of have to *stand* there while you two stand there laughing at me. (CONV)
> He *stood* alone in the empty hall. (FICT)

In contrast, *begin* is more commonly used in a non-physical sense, marking an aspectual process of 'beginning' relative to some other physical activity, event, or process, which is described in the following complement clause. For example:

> And then it *began* to get a bit darker. (CONV)
> I *began* to cry… (FICT)

Similarly strong use patterns distinguish other pairs of verbs with the same valency potentials. For example, the verb *try* has an even stronger preference for a following complement clause than *begin*. In contrast, the verb *meet* has a very strong preference for a following noun phrase as direct object, while the intransitive patterns and the pattern with a following complement clause are relatively rare.

In sum, corpus analysis here allows us to understand the different ways in which verbs are actually used. That is, although verbs often have the same potential of occurrence with different valency patterns, corpus analysis makes it clear that our actual use of such verbs is highly systematic, with each verb having its own preferred patterns, depending on its typical meanings and functions.

CONCLUSION

The present paper has illustrated the highly systematic patterns that structure our everyday use of linguistic features in speech and writing. In other studies, I have documented a related kind of pattern: the linguistic co-occurrence patterns that comprise the dimensions of variation among spoken and written registers (e.g., Biber, 1988, 1995). Both kinds of patterns operate below the level of conscious awareness and are usually not accessible to native intuitions. However, as the above analyses illustrate, these are extremely powerful patterns that correspond to major differences among sets of words, grammatical variants, lexico-grammatical associations, and registers.

Awareness of these patterns of use is obviously important for both teachers and students. This is not to say that frequency information can be mechanically translated into materials for instruction and assessment. For example, an additional consideration is the ease/difficulty of learning for particular features. However, it is also the case that we can no longer afford to ignore the typical patterns of use identified by quantitative corpus analysis. Instead, we can look forward to important gains for students as we

begin to develop materials that reflect the actual patterns of use in particular registers.

APPEDIX

get	say	go	know	have	think	see	want	come	mean	take	make	give	
9597	7005	7346	6845	5038	4556	3277	3033	2960	2472	1591	1259	1024	conv
1295	5818	1652	846	2838	706	1026	792	920	226	1663	1954	1073	news

V	V+A	V+NP	V+Comp Cls		
I	IA	T	TCLS		
26	55	19	0	stand	Conv
26	7	67	0	change	
9	7	84	0	meet	
15	5	5	75	begin	
14	0	29	57	try	
I	IA	T	TCLS		
30	56	4	10	stand	fict
33	14	53	0	change	
13	6	81	0	meet	
17	5	4	74	begin	
15	0	12	73	try	

I	IA	T	TCLS				
26	55	18	1	stand	news		
30	12	57	1	change			
2	16	82	0	meet			
27	30	17	26	begin			
5	0	11	84	try			
I	IA	T	TCLS				
38	56	4	2	stand	acad		
26	17	57	0	change			
7	12	80	1	meet			
16	30	14	40	begin			
7	0	5	88	try			
26	55	19	0	STAND - Conv	Conv	Conv	
26	55	18	1	STAND - News	news	News	
15	5	5	75	BEGIN - Conv		Conv	
27	30	17	26	BEGIN - News		News	
14	0	29	57	TRY - Conv		Conv	
5	0	11	84	TRY - News		News	
9	7	84	0	MEET - Conv		Conv	

2	16	82	0	MEET - News	News

	top 12	other lex
tot corpus		
Conversation	51200	73158
Fiction	32200	102367
News	18000	84240
Academic	8800	74528

simple	perfect	progressive	
119000	6300	7000	Conversation
110600	9759	5800	Fiction
78890	8500	4500	News
72000	4884	1500	Academic

present	past progressive	
5040	2020	Conversation
2160	3660	Fiction
3080	1400	News
1080	400	Academic

Verb + THAT-clause	"Extraposed" THAT-clause	Verb + TO-clause	"Extraposed" TO-clause	
6100	100	2500	100	Conversation
1400	600	2700	1450	Academic prose
V + Direct Object	V + Clause	V + Indirect Object + Clause		
5	5	80		TELL
30	55	5		PROMISE
2050	1250	760	450	1530
THINK	SAY	KNOW	GUESS	All other verbs
Classroom teaching	Classroom management	Textbooks	Syllabi, etc.	
8.1	9.2	7.2	6	Possibility modals
3.9	4.8	2.3	6.1	Necessity modals
12.3	22.9	4.5	16.7	Prediction modals

Figure 1: Distribution of the most common lexical verbs vs. other verbs, across registers (based on Biber et al, 1999, Figure 5.8)

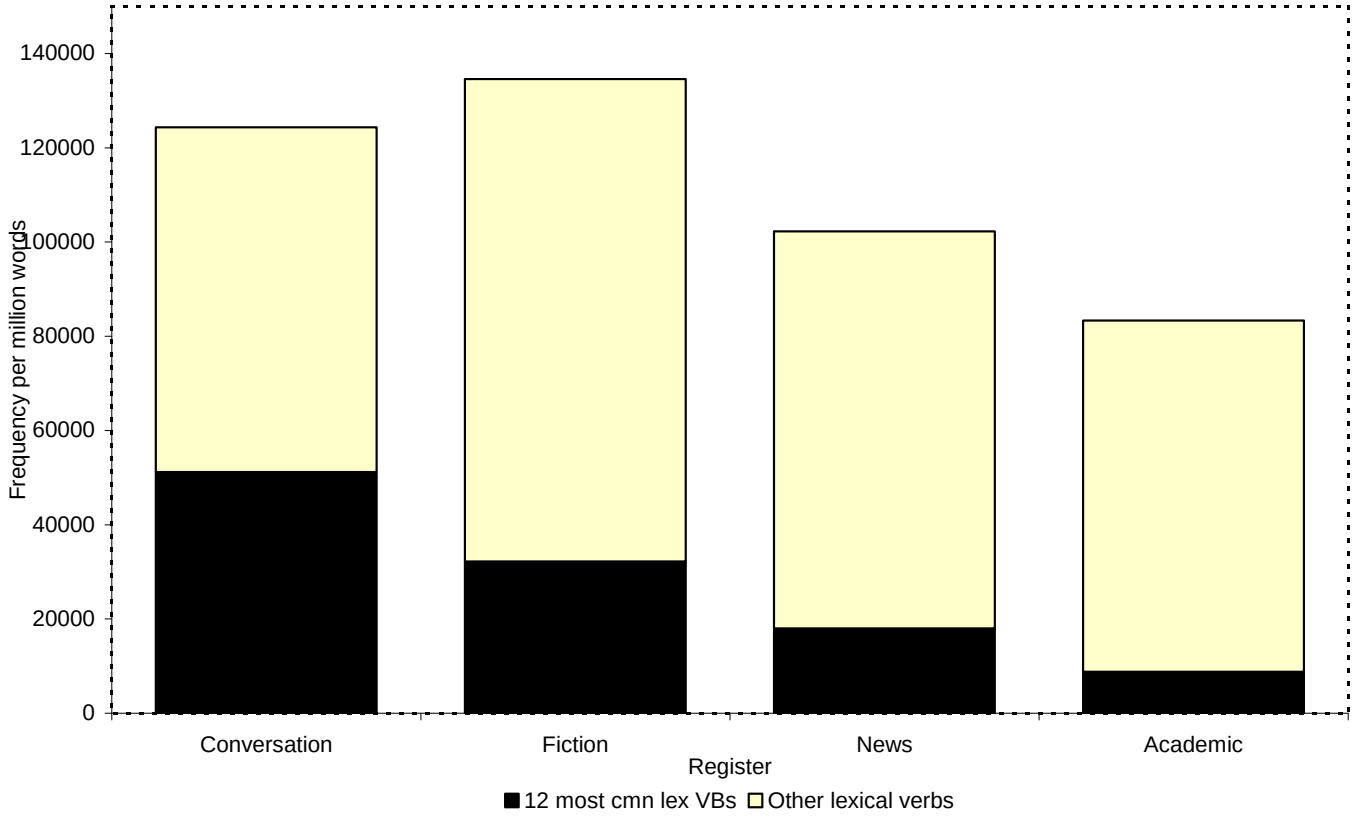


Figure 2: Frequencies in conversation of the
most common lexical verbs (based on Biber et al, 1999, Figure 5.9)

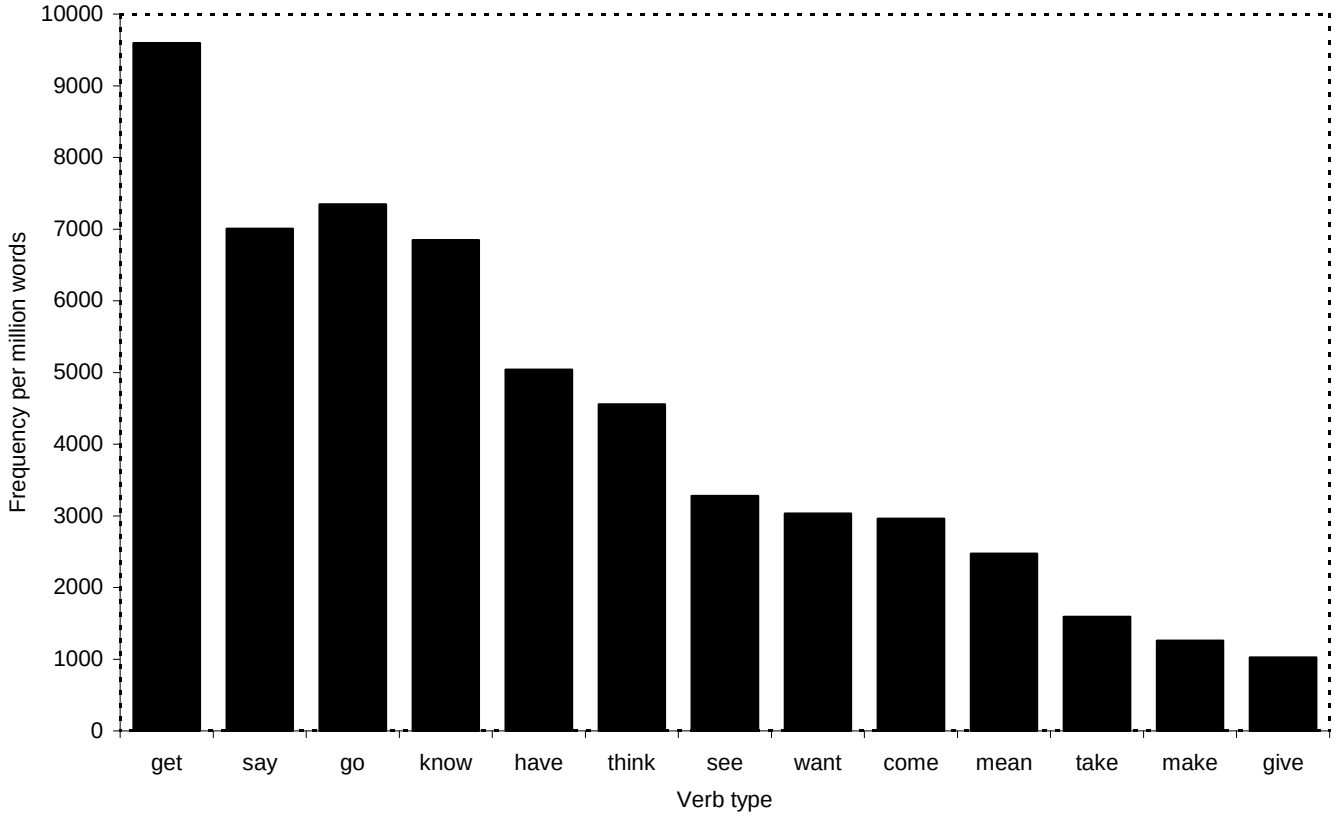


Figure 3: Frequencies in Newspapers of
most common lexical verbs (based on Biber et al, 1999, Figure 5.11)

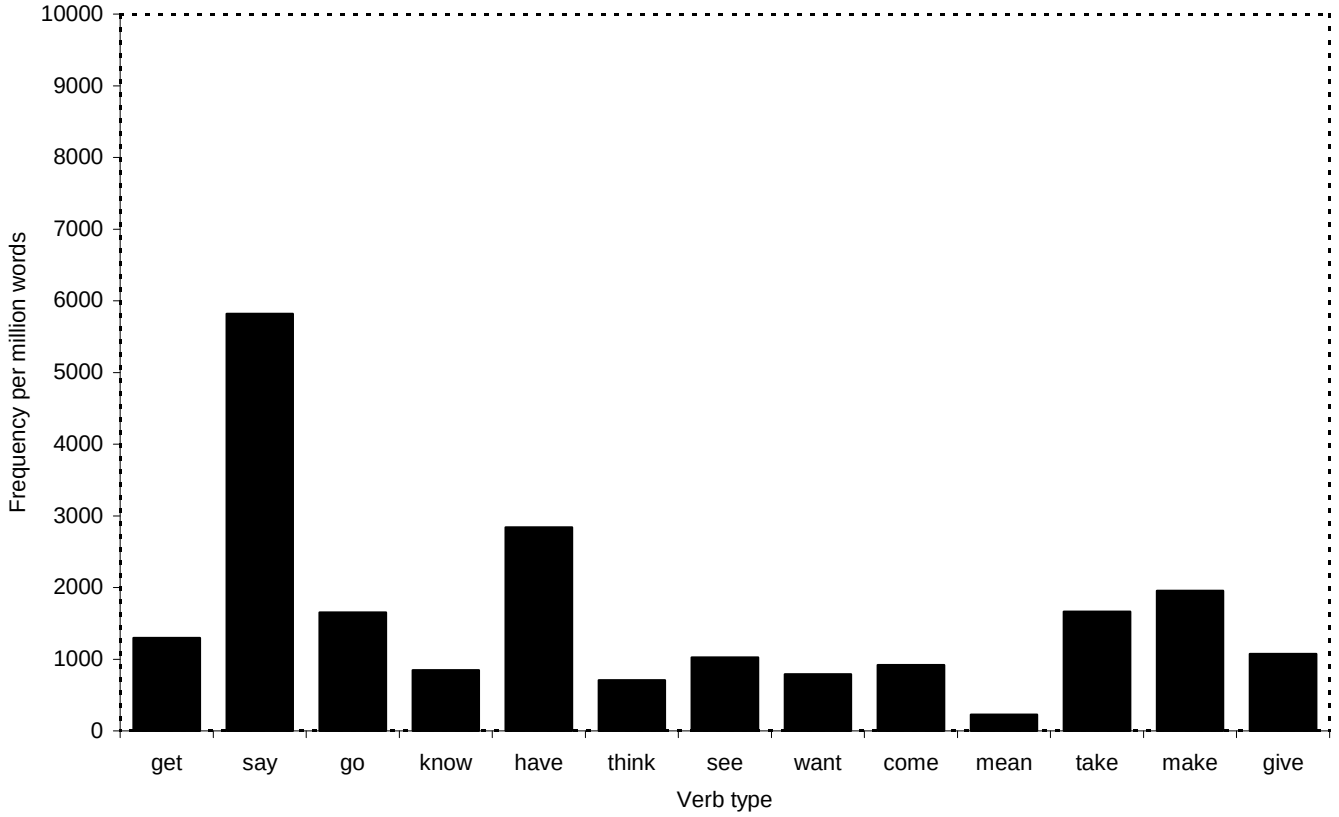


Figure 4: Frequency of present progressive and
past progressive in four registers (based on Biber et al, 1999, Figure 6.4)

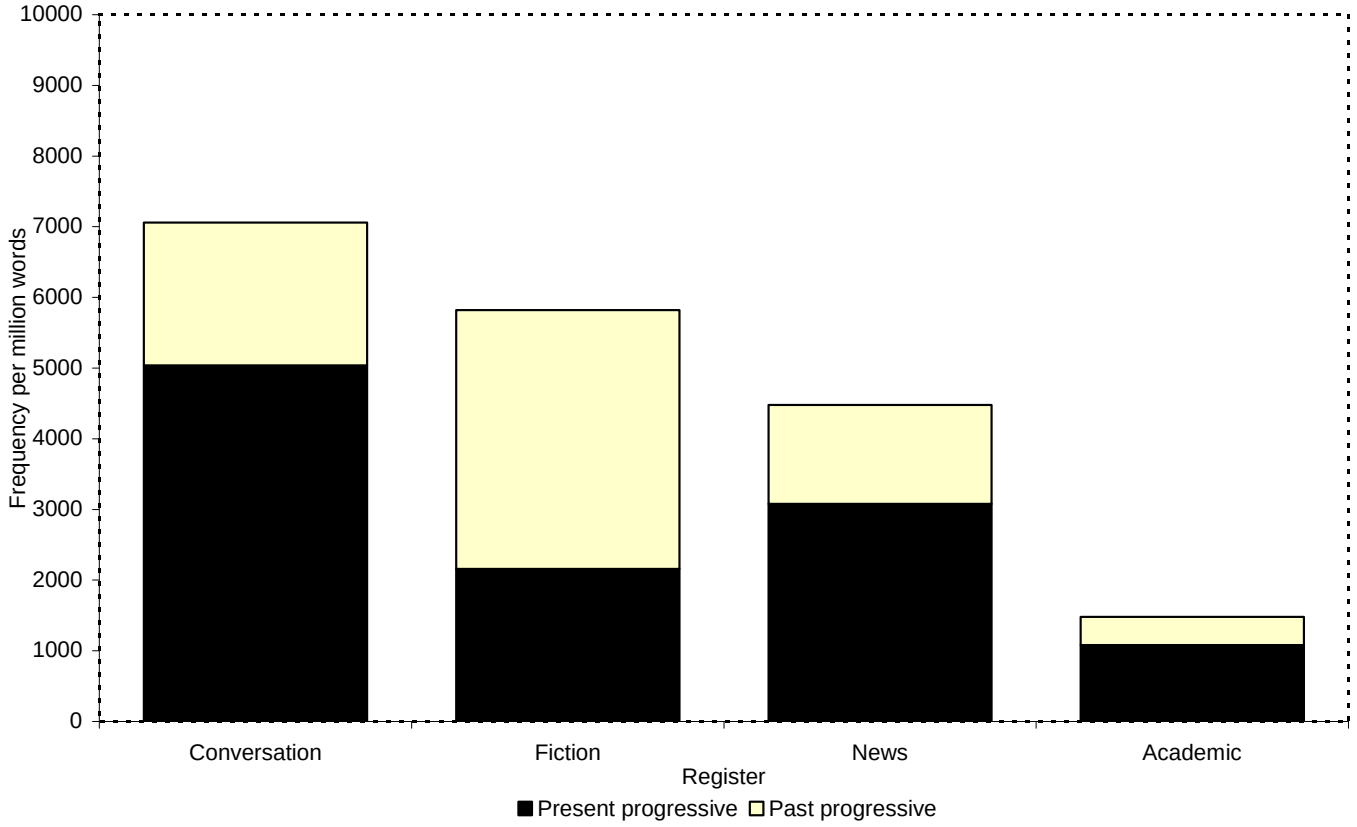
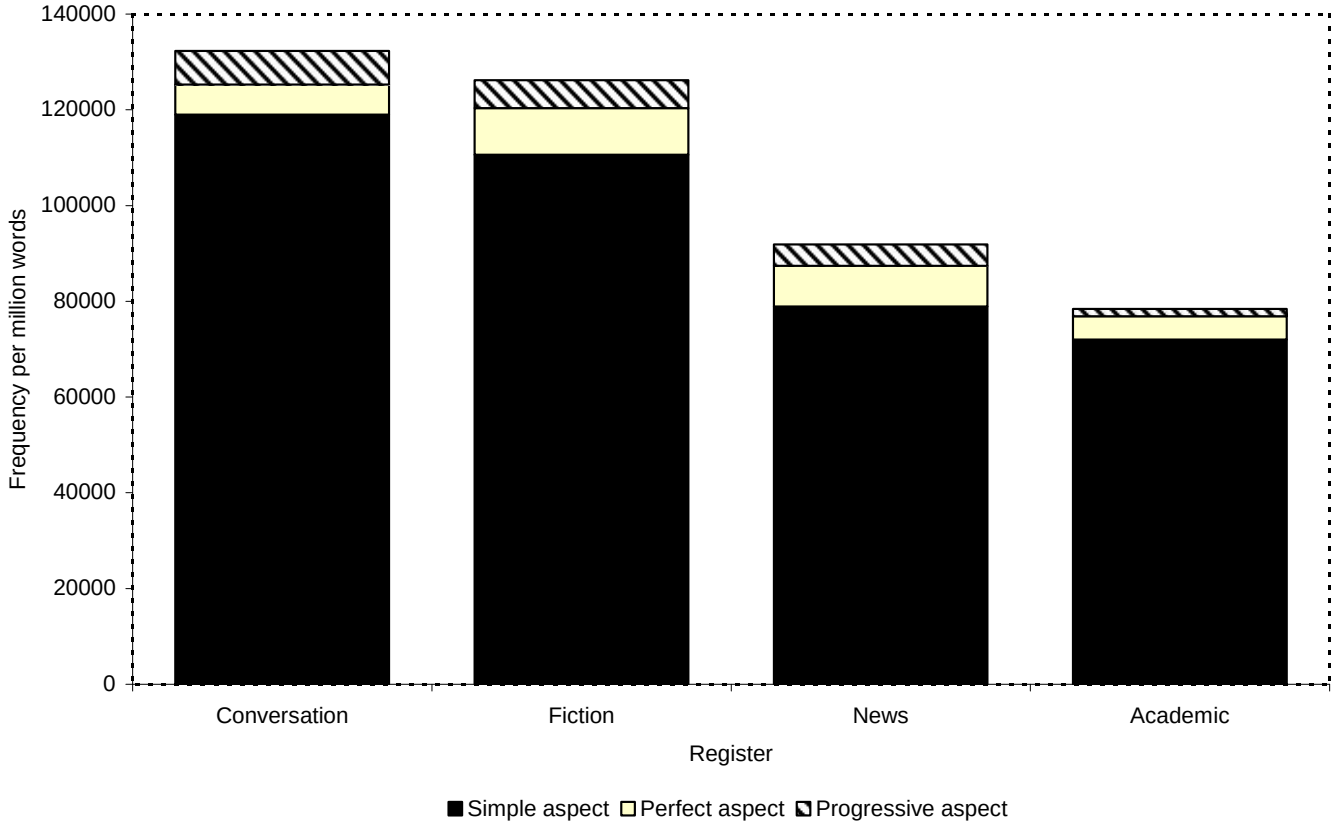


Figure 5: Frequency of simple, perfect, and progressive aspect in four registers (based on Biber et al, 1999, Figure 6.2)



REFERENCES

Aarts, B. & Meyer, C. (Eds.). 1995. *The verb in contemporary English: theory and description*. Cambridge: Cambridge University Press.

Aijmer, K. & Altenberg, B. (Eds.). 1991. *English corpus linguistics*. London: Longman.

Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. 1995. *Dimensions of register variation: a cross-linguistic perspective*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Reppen, R. 1998. *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. 1999. *Longman grammar of spoken and written English*. London: Longman.

Collins, P. 1991. *Cleft and pseudo-cleft constructions in English*. London: Routledge.

de Haan, P. 1989. *Postmodifying clauses in the English noun phrase: a corpus-based study*. Amsterdam: Rodopi.

Fox, B. A., & Thompson, S. A. 1990. A discourse explanation of the grammar of relative clauses in English conversation', *Language 66*, 297-316.

Geisler, C. 1995. *Relative infinitives in English*. Uppsala: Uppsala University.

Granger, S. 1983. *The* be + past participle *construction in spoken English with special emphasis on the passive*. Amsterdam: Elsevier Science Publishers.

Johansson, C. 1995. *The relativizers whose and of which in present-day English: description and theory*. Uppsala: Uppsala University.

Johansson, S., & A-B. Stenström (Eds.). 1991. *English computer corpora: selected papers and research guide*. Berlin: Mouton.

Mair, C. 1990. *Infinitival complement clauses in English*. New York: Cambridge University Press.

Meyer, C. 1992. *Apposition in contemporary English*. Cambridge: Cambridge University Press.

Myhill, J. 1995. Change and continuity in the functions of the American English modals, *Linguistics 33*, 157-211.

Myhill, J. 1997. *Should* and *Ought*: the rise of individually oriented modality in American English, *English language and linguistics 1*, 3-23.

Prince, E. F. 1978. A comparison of *Wh*-clefts and *It*-clefts in discourse, *Language 54*, 883-906.

Schiffrin, D. 1981. Tense variation in narrative, *Language 57*: 45 – 62.

Schiffrin, D. 1985a. Multiple constraints on discourse options: a quantitative analysis of causal sequences, *Discourse processes 8*, 281 – 303.

Schiffrin, D. 1985b. Conversational coherence: the role of *well*, *Language 61*, 640 – 667.

Schiffrin, D. 1987. *Discourse markers*. Cambridge: Cambridge University Press.

Sinclair, J. 1991. *Corpus, concordance and collocation*. Oxford: Oxford University Press.

Thompson, S. A. 1983. Grammar and discourse: the English detached participial clause. In F. Klein—Andreu (Ed.) *Discourse perspectives on syntax*. New York: Academic Press.

Thompson, S.A. 1985. Grammar and written discourse: initial vs. final purpose clauses in English, *Text 5*, 55 – 84.

Thompson, S. A., & Mulac, A. 1991a. The discourse conditions for the use of the complementizer *That* in conversational English, *Journal of Pragmatics 15*, 237-51.

Thompson, S.A., & Mulac, A. 1991b. A quantitative perspective on the grammaticization of epistemic parentheticals in English. In E.C. Traugott & B. Heine (Eds.). *Approaches to grammaticalization: volume II*. Amsterdam: John Benjamins.

Tottie, G. 1991. *Negation in English speech and writing: a study in variation*. San Diego: Academic Press.

Varantola, K. 1984. *On noun phrase structures in engineering English*. Turku: University of Turku.