English SIT, STAND, and LIE in small and large corpora

John Newman
Massey University

Sally Rice
University of Alberta

Abstract
We explore the usage of the cardinal English posture verbs *sit*, *stand*, and *lie* relying on a number of corpora of English with a view towards establishing quantitative and qualitative differences for these search items across relatively small and relatively large corpora. Frequencies of these verbs in various contexts were compared, as well as the range of correlates found in some selected constructions. Our results reveal a high degree of consistency between all the corpora for frequency of occurrence of the posture verbs. However, only the very large corpus revealed the full extent of collocational patterning exhibited by these verbs, which for most linguists is the primary reason for engaging in corpus-based investigation.

1 Introduction

The general context of the present study is a larger project which examines the syntagmatic behavior and pragmatics of posture verbs in English and other languages (Newman in press a, and the papers contained therein; especially Newman in press b; Newman and Rice 2001). The overall project is concerned in particular with the grammaticalization of 'sit', 'stand', and 'lie' verbs as tense and aspect markers and noun classifiers. English *sit*, *stand*, and *lie* have not undergone grammaticalizations of this sort. Nevertheless, we believe that patterns of collocation with these verbs in English reveal patterns which in some ways mirror the patterns of grammaticalization evident in some other languages (cf Heine et al 1991; Heine et al 1993; Kuteva 1999). Obviously, introspection and casual observation are of limited value in what they can and cannot yield in terms of insights into linguistic patterning over multiple lemma sets in multiple

contexts. Therefore, we have conducted searches for these verbs and their constructional and usage contexts in a variety of spoken and written English corpora, with the aim of ascertaining whether these posture verbs in English are beginning to show any of the pragmatic and grammatical tendencies that similar verbs are susceptible to in other languages.

One reason for using multiple corpora is to increase the number of sample contexts and to see whether spoken English tendencies are, as yet, beginning to manifest themselves in more formal registers. In addition, De Beaugrande (1996: 512) has argued that a very large corpus (such as the British National Corpus [henceforth BNC]) can yield results which have no direct counterpart in smaller corpora. We thought it would be an interesting exercise, therefore, to compare some of our findings based on the BNC (the main source for our research on posture verbs) with the results of identical searches in the smaller ICAME corpora as a way of further testing de Beaugrande's ideas.

2 The corpora

We made use of the eight corpora of written English listed in (1) and the three corpora based on transcriptions of spoken English listed in (2):

(1)	**BNC**	British National Corpus	100 million words	1970's-1990's
	LOB	Lancaster-Oslo/Bergen Corpus	1 million words	1960's
	FLOB	Freiburg-LOB	1 million words	1990's
	BROWN	Brown Corpus of American English	1 million words	1960's
	FROWN	Freiburg-Brown Corpus of American English	1 million words	1990's
	ACE	Australian Corpus of English	1 million words	1980's
	WC	Wellington Corpus of Written New Zealand English	1 million words	1986-1990
	KOL	The Kolhapur Corpus of Indian English	0.9 million words	1978

(2)				
	LLC	London-Lund Corpus	0.5 million words	1950's-1980's
	COLT	Corpus of London Teenage Language	0.5 million words	1993
	WSC	Wellington Corpus of Spoken New Zealand English	1 million words	1988-1994

Of the written corpora we used in our study, the BNC was by far the most impressive in terms of its size, and it was the corpus that we relied on most. The BNC consists of 90 percent written sub-corpora and ten percent spoken. The remaining written corpora, distributed as part of the ICAME package of corpora, are comparable in so far as they all follow roughly similar principles in their selection of texts, with the BROWN and LOB corpora providing the early models. The spoken corpora, also taken from the ICAME set, are more diverse in their selection of material, as suggested by the titles of the corpora.

While it was of some interest to us to compare search results across various corpora, there was no strong expectation or requirement that the frequency results obtained from the BNC should correspond exactly with results obtained from identical searches in the remaining written corpora. As de Beaugrande (1996: 512) has pointed out, a very large corpus (such as the BNC) can yield results which have no direct counterpart in the smaller corpora such as the one-million word corpora and less. More specifically, there can be qualitatively different kinds of results obtainable from the very large corpora. For example, numerous items in the largest corpus do not appear in the smaller corpora. More informed judgments about relative frequency are also possible with large corpora. Very large corpora naturally display finer degrees and differentiations of detail, which is especially useful when looking for collocational patterning of items, as we were. Furthermore, there can be profound sampling differences associated with written and spoken corpora as well as small and large corpora, both with respect to depth of the discourse sampled and to the number of different sources, all of which further exacerbates natural differences attributable to modality and register.

It was a matter of some interest to us, nevertheless, to establish by means of common statistical techniques the extent to which various search results from the different corpora could be said to 'correlate' with each other. While we accept de Beaugrande's points about an essential qualitative difference between the BNC and the other corpora, we were still interested to know whether the search results were relatively uniform across all the corpora. Where a result has

a high consistency across all the corpora, large or small, then that result seems especially worthy of attention. Equally noteworthy are results which are only evident in, say, the largest corpus, the BNC.

3 Comparisons of frequencies

3.1 Key constructions

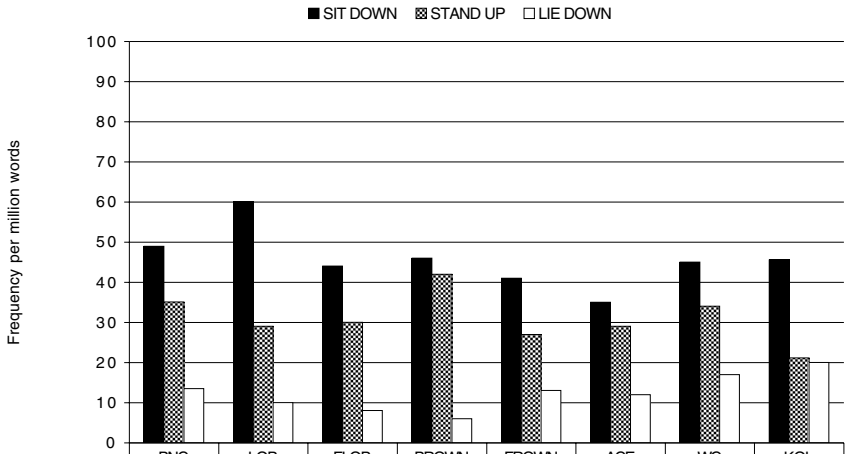
Our first focus was on the set of results from the following search strings:

(3) SIT DOWN = sit/sits/sitting/sat down
STAND UP = stand/stands/standing/stood up
LIE DOWN = lie/lies/lying/lay/lain down

These strings were chosen as a way of probing the use of the lexemes SIT, STAND and LIE in ways which are closely associated with maintenance or assumption of posture, while largely excluding other senses. The intransitive posture verb *lie* and the transitive posture verb *lay* are prone to some confusion in English (eg speakers will use both intransitively, as in *she lay down to sleep* and *she laid down to sleep*). This conflation thus adds to the difficulty of identifying by any simple procedure only the intransitive uses. Furthermore, the addition of *down* to *lie/lies/lying* largely eliminates any possibility of the 'tell a lie' meaning of *lie*. There is, of course, still a wide range of non-postural meanings associated with these forms (*she was stood up*, *stand up for one's rights*, *stand up to someone*, *stand up to scrutiny*, *to lay down the law* etc). Further querying of the corpora and line-by-line inspection would have been necessary to narrow down the range of meanings in this way and this was not the highest priority for us. We were primarily interested in exploring the effectiveness of simple computerized searches, rather than carrying out manual inspections of the corpora.

The corpora were searched exhaustively rather than by random sampling from a set number of words or lines. Since the corpora vary in size, we normalized the results in terms of occurrence per million words. This has the advantage of facilitating comparison of frequencies across the corpora, but it does disregard the huge differences in magnitude of the BNC database vis-à-vis the remaining corpora. This difference in magnitude is very relevant to understanding some of the differences in collocational patterns evident in the corpora, a point to which we return in section 4. Figure 1 is an overview of the results in

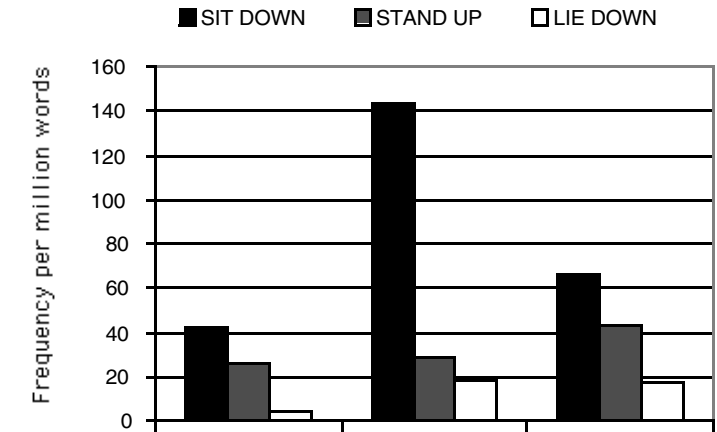
the written corpora shown in graph form, with the data table showing the normalized figures representing the number of occurrences per million words. As noted above, the BNC consists of both written and spoken sub-corpora; however, we included the BNC results in Figure 1 since it represents overwhelmingly written texts. Figure 2 shows the results for the spoken corpora.



	BNC	LOB	FLOB	BROWN	FROWN	ACE	WC	KOL
■ SIT DOWN	48,93	60	44	46	41	35	45	45,5
⊠ STAND UP	35,09	29	30	42	27	29	34	21,1
□ LIE DOWN	13,5	10	8	6	13	12	17	20

Figure 1: Normalized frequency of posture verbs in written corpora (per millions words)

A number of patterns are observable in these results. First and foremost is the fact that in every corpus, from the 100-million-word written BNC to the much smaller half-million-word spoken corpora like the COLT and LLC, the same overall order of frequency obtains: SIT DOWN > STAND UP > LIE DOWN. Being the largest corpus by far, the BNC results are especially significant and the results from this corpus show a clear gradation and separation in frequency.



	LLC	COLT	WSC
SIT DOWN	42	144	66
STAND UP	26	28	43
LIE DOWN	4	18	17

Figure 2: Normalized frequency of posture verbs in spoken corpora (per million words)

These searches, done purely on the basis of matching for forms, produced results comparable to what is known about the relative frequency of the posture senses based on the semantically tagged files of the BROWN corpus. As is well known, most of the BROWN corpus has been manually tagged in a way which identifies, amongst other things, sub-meanings of verb lexemes in accordance with the requirements of WordNet (Fellbaum 1998). In the case of the posture verbs we are interested in, WordNet differentiates posture senses from other senses (eg the 'tell a lie' sense of *lie*), as well differentiating the particular sub-meanings of posture (eg 'assume a posture' vs 'maintain a posture'). The BROWN corpus, tagged in this way, yields the usage totals given in Table 1 for the lexemes SIT, STAND, and LIE. The top row gives the frequency of occurrence for these verbs when used statively, without any accompanying verb particle such as *up*, *down*, *back* etc; the lower row gives occurrences for when these verbs are used dynamically, with the verb particles indicated. The numbers refer to tokens of the lexemes, ie all the inflected forms of the verb with the specific meaning indicated. Lexemes are written in small caps.

Table 1: Occurrences of posture verbs with specific senses in the BROWN corpus

SIT **47** 'be seated, be in a sitting position'	STAND **43** 'be standing, be upright'	LIE **35** 'be lying, be prostrate/prone'
SIT DOWN **10** 'take a seat'	STAND UP **8** 'arise, get up'	LIE DOWN **1** 'assume a reclining position'

The search results from BROWN in Figure 1 are slightly different in actual numbers from those in Table 1 because the search in Figure 1 returned STAND UP forms with meanings such as 'be standing, be upright', 'refuse to back down, remain solid under criticism or attack', 'defend against attack or criticism'. With the aid of WordNet (and its accompanying concordance utility Semcor) one is able to achieve results specifically for the meaning 'arise, get up' and these results are reflected in Table 1. It is interesting to note, however, that the coarser probing of the corpora using form-only search strings did in fact yield comparable results to what was found with the very fine-grained searches based on the semantically tagged text.

A comparison of the spoken and written corpora reveals further differences in frequencies, namely a generally higher frequency of occurrence in spoken English compared with written English for expressions containing posture usages of these predicates. Consider SIT DOWN. In the written corpora it usually occurs less than 20 times per million words, whereas in the spoken corpora it occurs more than 20 times per million words. A preference for SIT DOWN in spoken language is most pronounced in COLT, a corpus of entirely informal conversation. This fact appears to reflect a preference for basic vocabulary items (such as the posture verbs we are researching) in conversation, especially in informal speech.

3.2 Other constructions

In order to broaden our search, we obtained frequencies of occurrence for an assortment of posture lexeme + {preposition/adverb/ particle} strings. The 42 search strings are listed in Table 2. The choice of preposition, adverb, or particle

was influenced by our interest in posture-related meanings of *sit*, *stand*, and *lie* words and we felt that these particular forms tended to occur with posture meanings, as opposed to meanings such as 'tell a lie' or highly figurative meanings. The frequencies obtained in this way, normalized as number of occurrences per million words, are given in Appendix 1 (written corpora) and Appendix 2 (spoken corpora). Note that it is not always the SIT expression which occurs with highest frequency of the three cardinal posture verbs. For example, STAND BY occurs with greater frequency than either SIT BY or LIE BY. One reason for this is that, in a number of these expressions, like STAND BY, we do indeed find figurative extensions of the posture verbs, where the extensions occur only with one of the posture verbs.

We then carried out a series of pairwise comparisons of the results from the written corpora across the 42 search strings we investigated in order to establish the degree of consistency in the results. We did the same for the three spoken corpora. Using Fisher's r to z function, a standard test for correlations, a p-value of less than 0.0001 was returned for every one of these pairwise correlations, as shown in Table 3.

Table 2: Additional search strings containing the posture verbs

SIT	STAND	LIE
+ *down*	+ *up*	+ *down*
+ *here/there*	+ *here/there*	+ *here/there*
+ *in*	+ *in*	+ *in*
+ *on*	+ *on*	+ *on*
+ *at*	+ *at*	+ *at*
+ *with*	+ *with*	+ *with*
+ *by*	+ *by*	+ *by*
+ *next to*	+ *next to*	+ *next to*
+ *beside*	+ *beside*	+ *beside*
+ *behind*	+ *behind*	+ *behind*
+ *over*	+ *over*	+ *over*
+ *above*	+ *above*	+ *above*
+ *under*	+ *under*	+ *under*
+ *underneath*	+ *underneath*	+ *underneath*

Table 3: Correlations between frequencies of posture expressions in Table 2 across the corpora

Fisher's r to z		
	Correlation	P-Value
BNC, LOB	0.932	<.0001
BNC, FLOB	0.962	<.0001
BNC, BROWN	0.959	<.0001
BNC, FROWN	0.924	<.0001
BNC, ACE	0.949	<.0001
BNC, WC	0.928	<.0001
BNC, KOL	0.878	<.0001
LOB, FLOB	0.956	<.0001
LOB, BROWN	0.927	<.0001
LOB, FROWN	0.872	<.0001
LOB, ACE	0.853	<.0001
LOB, WC	0.873	<.0001
LOB, KOL	0.869	<.0001
FLOB, BROWN	0.943	<.0001
FLOB, FROWN	0.916	<.0001
FLOB, ACE	0.909	<.0001
FLOB, WC	0.919	<.0001
FLOB, KOL	0.839	<.0001
BROWN, FROWN	0.911	<.0001
BROWN, ACE	0.906	<.0001
BROWN, WC	0.873	<.0001
BROWN, KOL	0.872	<.0001
FROWN, ACE	0.939	<.0001

FROWN, WC	0.891	<.0001
FROWN, KOL	0.861	<.0001
ACE, WC	0.942	<.0001
ACE, KOL	0.853	<.0001
WC, KOL	0.848	<.0001

Fisher's r to z		
	Correlation	P-Value
LLC, COLT	0.839	<.0001
LLC, WSC	0.920	<.0001
COLT, WSC	0.930	<.0001

While the very low p-values appear impressive as significantly greater than chance correlations, these results need to be qualified. The written corpora, apart from the BNC, are not completely independent in their design and sampling methodology. Indeed, they form a natural set in the sense that the BROWN and LOB corpora were models for the subsequent corpora. They were designed to be comparable in terms of their coverage of genres and methods of collection of texts. The high correlation between the frequencies of the posture verbs in the ICAME written corpora is therefore not surprising. More significant is the high correlation between these corpora and the BNC and the high correlation between these corpora and the spoken corpora of ICAME. Neither the BNC nor the spoken corpora were designed or collected in ways directly comparable to LOB or BROWN. The spoken corpora in particular represent quite diverse kinds of material, and finding relatively high correlations that are statistically significant at the $p < .0001$ level is more relevant for our purposes.

4 Comparison of collocates

In a subsequent round of searches, we investigated selected collocational patterns with the posture verbs in the hope of revealing more qualitative information about the actual usage (as opposed to usage potential) of these verbs. In particular, we were interested in determining the range of verbal predicates

which had a close association with the posture verbs. As a way of operationalizing 'close association', we searched for instances of the posture verbs occurring as the first conjunct in a conjunction of verbs with *and*. We distinguished two different searches along these lines:

(4) a. {*sitting/standing/lying*} *and...V-ing*

b. {*sat down/stood up/lay down*} *and...V-ed*

The *V-ing* forms in (4a) could be present participles or gerunds while the *V-ed* forms in (4b) involved either past tense or past participial forms. The (a) cases correspond semantically to simultaneous conjunction, ie the conjoined expression refers to a posture and the actions/states of the clause participant while in that posture. The (b) forms correspond to a consecutive conjunction, referring typically to the action of entering a posture and the actions/states which the person(s) engaged in afterwards. Because we anticipated relatively low frequencies of occurrence for these rather specialized usage contexts, we initially queried only the BNC for these expressions. However, selected databases in the ICAME corpora also revealed comparable results, though at a much reduced rate of return.

4.1 Simultaneous conjuncts

The search results from the BNC for instances of simultaneous conjuncts are shown in Table 4:

Table 4: Posture verbs in simultaneous conjunction in the BNC

sitting and…V-ing		*standing and…V-ing*		*lying and…V-ing*	
Total	**58**	**Total**	**46**	**Total**	**3**
V-ing tokens >1		*V-ing* tokens >1		*V-ing* tokens = 1	
watching	8	*looking*	3	*talking*	1
listening	6	*staring*	3	*waiting*	1
waiting	5	*talking*	3	*sitting*	1
looking	4	*balancing*	2		
reading	4	*bending*	2		

thinking	4	*sitting*	2	
talking	3			
saying	2			
staring	2			

Once again, the relative order of collocate frequency is the same as found in the searches above: *sitting* (58) > *standing* (46) > *lying* (3). *Lying*, as the first conjunct, occurs with considerably less frequency than the other two posture verbs. Lying is not normally a posture we maintain while simultaneously carrying out other activities and the extreme infrequency of the *lying and ...* construction, as well as the absence of any recurring combinations, reflects this. In the case of *sitting* and *standing*, there are clear and recurring patterns of conjuncts; ie we find multiple instances of the same combinations. We have listed the *–ing* forms which occur as the second conjunct in these recurring combinations. These collocate verbs are by no means random. Rather, they can be experientially motivated as reflecting recognizable states or activities which conventionallly accompany the different postures. Reading and thinking, for example, are activities which may take time to complete and are normally done in a comfortable position which can be maintained. Reflecting this, *reading* and *thinking* occur as the second conjunct of *sitting*, but not *standing* or *lying*. More generally, *sitting* occurs preferentially with verbs which have a perceptual, cognitive, or otherwise ideational and reflective meaning. Standing, as a posture, requires the highest degree of sensorimotor control compared with the other postures considered here and, correlated with this, we find that *balancing* and *bending* form recurring patterns in combination with *standing* (consistent, incidentally, with the findings of Gibbs et al 1993) but not *sitting* or *lying*. The BNC results, therefore, reveal patterns in these simultaneous conjuncts with the posture verbs which are significant in the way they pick out experientially regular co-occurrences of postures and other collateral activities or states. These patterns of collocation are particularly interesting in the context of research into grammaticalization processes affecting posture verbs, especially the aspectualization of posture—a phenomenon which we elaborate on elsewhere (cf Newman and Rice 2001).

Searching the same strings in the ICAME corpora yielded the results in Table 5. Note, there were no multiple instances of identical collocates in any single one of these much smaller corpora.

Table 5: Posture verbs in simultaneous conjunction in selected ICAME corpora

	sitting and…V-ing	*standing and…V-ing*	*lying and…V-ing*
Total	**7**	**2**	**0**
LOB	——	——	——
FLOB	*worrying*	*watching*	——
BROWN	*staring*	——	——
FROWN	*talking*	——	——
ACE	——	——	——
WC	——	——	——
KOL	*staring*	*giving (orations)*	——
LLC	——	——	——
COLT	——	——	——
WSC	*waiting and talking* *working*	——	——

Taken individually, the ICAME corpora do not yield consistent results on these searches. In some of the corpora, none of the posture verbs is found in a simultaneous conjunction, making it impossible to ascertain any differential behavior between the three posture verbs in terms of frequency of occurrence. In other cases, there is only the merest hint of different collocational patterns. BROWN, FROWN, and WSC, for example, show just one instance of simultaneous conjunction with *sitting and*, compared with no instances at all for *standing and* and *lying and*. This shows, arguably, a preference for the *sit* verb in such constructions, consistent with the BNC (and the preponderance of SIT verbs as durative and progressive markers cross-linguistically), but it fails to discriminate between *standing* and *lying*.

If we take the ICAME corpora collectively, then we see an emergent pattern of relative frequency of the posture verbs in simultaneous conjunction, namely *sitting* (7) > *standing* > (2) > *lying* (0). Even so, the absolute numbers involved are exceedingly small and one can not be very confident about the results. The only recurring patterns evident in the (combined) ICAME sub-corpora involved

sitting and talking and *sitting and staring*, both of which were manifested in two instances.

4.2 Consecutive conjuncts

We searched the BNC for instances of any consecutive conjuncts formed by a conjunction of the *-ed* form of the posture verb followed by *and* V-*ed*, where the *–ed* refers to either the past tense of past participial form of some following verb. The results are summarized in Table 6:

Table 6: Posture verbs in consecutive conjunction in the BNC

sat down and…V-ed		*stood up and…V-ed*		*lay down and…V-ed*	
Total	**261**	**Total**	**477**	**Total**	**20**
collocate *V-ed* tokens > 2		collocate *V-ed* tokens > 2		collocate *V-ed* tokens > 1	
watched	6	*walked*	30	*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	

The relative frequency of the three posture verbs is noticeably different from that found with the simultaneous conjuncts. In the case of the consecutive conjuncts, the order of frequency is *stood up* > *sat down* > *lay down*. Note that this is different also to the relative frequency of the posture verbs with the 'assumption of posture' senses identified in the semantically tagged BROWN corpus and cited above in Table 1. The key to this difference lies in the sequence of the posture verb followed by *and*, leading to an interpretation involving the assumption of a posture followed by a further activity or state. Thus, even though SIT DOWN in the 'sit down, take a seat' is more common than STAND UP in the 'arise, get up' sense in the BROWN corpus, this does not imply that these verbs occur with the same relative frequency in conjoined contexts. And indeed, this is not the case in the BNC (or various other corpora, as we will see below). The experiential reality underlying this fact is that we stand up in order to carry out more

numerous and more varied collateral activities than we do when we sit or lie down.

It is not just the different relative frequency which is noteworthy about the consecutive conjunction. The semantic content of the three sets of verbs in Table 6 is also remarkable. What is striking about the *stood up and V-ed* set of verbs is the occurrence of verbs of motion and physical activity, verbs which are absent from the other two sets. Such verbs include: *walked, went, moved, stretched, put, came, put, left, reached, wandered*, *strode*. These refer to actions one proceeds to carry out after having stood up. The BNC results reflect the naturalness of these sequential scenarios and the relative unnaturalness of such activities as a result of sitting down or lying down. It is reminiscent of the fact that *standing*, but not the other posture verbs, collocates with *balancing* and *bending* as simultaneous conjuncts, as observed above. There are no such collocates with *sat down and V-ed* which occurs with an assortment of verbs of a more introspective and temporally extended nature, such as *watched*, *thought*, *waited*, *rested*. Almost all the recurring patterns with *lay down and V-ed* involve a verb referring to sleeping. These different classes of verbs collocating with the posture verbs reveal familiar experiential practice. As with the simultaneous conjuncts, the collocational patterns are of some interest to grammaticalization studies. Specifically, it is the STAND verb which is more likely to take on an inceptive sense, eg Old High German *stantan* 'stand; begin' (Schützeichel 1969: 182). Consider also Middle Egyptian *h* 'stand up, rise', which, when used with a past tense marker, forms an auxiliary with the meaning 'thereupon' (Gardiner 1957: 391-392; Heine et al 1993: 206). These developments are completely in keeping with the tendency in English for STAND (more so than SIT or LIE) to occur in a consecutive conjunction with verbs which represent the movement or activity following the change into a standing position (cf Newman and Rice 2001).

The same search strings in selected ICAME corpora revealed the patterns in Table 7:

Table 7: Posture verbs in consecutive conjunction in selected ICAME corpora (for which there are one or more instances; multiple instances are indicated in parentheses)

	sat down and…V-ed	*stood up and…V-ed*	*lay down and…V-ed*
Total	**19**	**38**	**2**
LOB	*studied* *ran his finger down* *offered [a cigarette]*	*walked* (2) *shouted* *took down [her bag]* *went into the living room*	——
FLOB	*crossed her legs*	*walked* (2) *gave [a smile]* *looked at his watch* *pushed his plate away* *tucked the book*	——
BROWN	*played two slots at once* *began [coffee party]*	*thanked* *began pacing* *added to the uproar* *beckoned* *demanded water* *nodded farewell* *went [to the windows]* *touched the girl's arm*	——
FROWN	*wrote*	*uncorked a right [punch]*	——

ACE	*had Morning Tea together*	*sighed* *yelled* *picked up [cups]* *raised her superior nose tossed his cigarette butt*	——
WC	*talked* *read* *made a hand-written copy*	*turned to the others* *held them against her* *began shouting* *stepped back a pace* *glared at her mother*	*gazed at the sky and dared not move*
KOL	*thought* *began to play*	*said* (2) *groped my way to [room] offered her seat in a bus*	——
LLC	——	——	——
COLT	*read* *visualized it*	——	——
WSC	*done a taste test* *deliberately thought* *picked up a crystal glass* *have a yarn*	*said* (2) *left* *made [a dick of herself]*	——

The collective totals for each posture verb present a clear pattern in terms of relative frequency, comparable to what we find in the BNC, ie *stood up* > *sat down* > *lay down*. For the majority of the ICAME corpora, the results of the searches show also manifest this relative frequency. Note, however, that FROWN in the

written corpora and LLC and COLT in the spoken corpora fail to discriminate the three verbs in this way.

As in the BNC, *stood up* also collocates with verbs referring to activities which involve motion and, particularly, motion away from the standing position: *walked*, *began pacing*, *went [to the window], stepped back a pace*, *left*. *Sat down*, too, occurs with verbs similar to those found in the BNC: *studied, read, thought,* etc. In the case of *lay down,* however, there is not even evidence of a collocation with verbs of sleeping. While we can therefore discern some similarity in these collocation patterns comparing the ICAME corpora and the BNC, it is really only in the totality of the ICAME corpora that these similarities appear. Individual corpora prove unreliable sources. FLOB, for example, includes examples of motion verbs with *stood up*, but lacks any examples of thinking or cognitive verbs with *sat down*; COLT includes examples of thinking and mental activity with *sat down*, but lacks examples of motion verbs with *stood up*; and a consecutive conjunct with *lay down* is absent from all but one of the ICAME corpora.

5 Conclusion

The design and construction of the BNC and ICAME corpora point to certain advantages and disadvantages which are obvious even without performing detailed interrogation of the corpora. The BNC is superior, of course, in terms of database size, but it is restricted to material that is almost entirely British. The ICAME corpora are small by comparison but have the advantage of coverage in their inclusion of different varieties of English (British, US, Indian, Australian, etc). These are properties of the various corpora which are known in advance. These general properties of the corpora, however, can not predict what one will find concerning frequency of and collocational patterning with particular lexical items.

By putting the various corpora through their paces, we were able to undertake a controlled parametric exploration of morphosyntactic, lexical, and pragmatic similarities and differences among members of a broad lexical category (posture verbs). Such an investigation is of some theoretical interest in typological, cognitive linguistic, and grammaticalization circles in linguistics, all of which are keenly interested in language variation and change and the motivations behind any such dynamic phenomena in language. The verbs we selected for the present study are drawn from the basic, everyday vocabulary of English (the most typical input sources of grammaticalization) and occur in all the corpora we have referred to. The consistency in the relative frequency of various

forms of the posture verbs *sit*, *stand*, and *lie* across the corpora is a striking quantitative result. Any one of the (relatively small) ICAME corpora allows one to obtain this result. Moving to the (relatively large) BNC does not alter this relative frequency; it only provides stronger confirmation of the same result. From this point of view, each of the selected ICAME corpora is as enlightening as the BNC. That is, even half a million words of text or speech (as in COLT and LLC) is sufficient to establish this result. This is not something one could have predicted just from a consideration of the general nature of the corpora.

Collocational patterns and their frequencies can be harder to establish than simple frequencies of lexical items (though our lexical items included some simple collocations such as *sit down*, *stand up,* etc), since the raw numbers of collocations of an item will always be less than that of its overall lemma or lexeme frequency. In the case of posture verbs in simultaneous and consecutive conjunction, the collocational patterns indeed involved a substantial reduction in frequency compared with the posture verbs alone. A number of collocational patterns well represented in the BNC are only barely in evidence in the totality of the ICAME corpora, while individual ICAME corpora may not reveal some of the targeted patterns at all. These results are of interest as we strive to understand the difference between small (<1 million word) corpora and large (>100 million word) corpora and establish the kinds of information retrievable from each. The collocations we have explored here were evident in the large corpus we used but not really apparent in any of the small corpora. Results like this help to exemplify and refine further de Beaugrande's (1996) claims about the nature of the differences between small and large corpora.

References

De Beaugrande, Robert. 1996. The 'pragmatics' of doing language science: The 'warrant' for large-corpus linguistics. *Journal of Pragmatics* 25: 503–535.

Fellbaum, Christiane (ed). 1998. *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press.

Gardiner, Alan H. 1957. *Egyptian grammar: Being an introduction to the study of hieroglyphs*. 3rd Edition. Oxford: Oxford University Press.

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

Heine, Bernd, Ulrike Claudi, and Friederike Hünnemeyer. 1991. *Grammaticalization: A conceptual framework*. Chicago and London: The University of Chicago Press.

Heine, Bernd, Tom Güldemann, Christa Kilian-Hatz, Donald A. Lessau, Heinz Roberg, Mathias Schladt, and Thomas Stolz. 1993. *Conceptual shift: A lexicon of grammaticalization processes in African languages*. Köln: Institut für Afrikanistisk, Universität zu Köln.

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

Newman, John. In press a. *The linguistics of sitting, standing, and lying*. Amsterdam and Philadelphia: John Benjamins.

Newman, John. In press b. An overview of 'sit', 'stand', and 'lie': A cross-linguistic perspective. In J. Newman (ed). *The linguistics of sitting, standing, and lying*. Amsterdam and Philadelphia: John Benjamins.

Newman, John and Sally Rice. 2001. English SIT, STAND, and LIE: Patterns of usage and their experiential motivations. Paper presented at The Seventh International Cognitive Linguistics Conference, Santa Barbara, CA: 22–27 July 2001.

Schützeichel, Rudolf. 1969. *Althochdeutsches Wörterbuch*. Tübingen: Max Niemeyer Verlag.

Appendix 1. Normalized frequency (per million words) of selected posture expressions in written corpora. Verbal lexemes, not just base forms, were used in the searches.

WRITTEN	BNC	LOB	FLOB	BROWN	FROWN	ACE	WC	KOL
SIT HERE/THERE	22.48	20	14	15	14	7	12	20
STAND HERE/THERE	16.94	20	12	27	13	11	22	21.11
LIE HERE/THERE	6.53	10	14	9	4	3	14	3.33
SIT IN	38.13	40	37	38	63	48	61	50
STAND IN	36.1	53	31	48	46	25	28	51.11
LIE IN	41.71	72	46	49	38	36	57	75.55
SIT ON	39.19	30	22	31	30	38	60	57.77
STAND ON	23.7	30	25	37	29	18	28	36.66
LIE ON	21.58	46	26	18	22	12	44	21.11
SIT AT	13.71	11	18	11	15	9	23	21.11
STAND AT	18.94	18	17	21	19	17	31	17.77
LIE AT	5.31	9	7	1	5	3	0	4.44
SIT WITH	8.07	13	4	7	4	7	5	11.11
STAND WITH	5.92	10	2	10	10	2	3	48.88
LIE WITH	5.75	15	5	7	2	3	7	8.88
SIT BY	3.95	3	4	5	5	2	5	5.55
STAND BY	12.28	13	12	11	10	6	17	15.55
LIE BY	0.57	0	0	1	1	0	1	1.11
SIT NEXT TO	3.15	1	4	0	2	4	7	3.33
STAND NEXT TO	1.23	2	2	3	5	1	1	4.44
LIE NEXT TO	0.46	1	1	1	0	0	1	0

SIT BESIDE	3.85	3	3	6	3	3	10	8.88
STAND BESIDE	3.64	3	2	4	8	8	1	1.11
LIE BESIDE	1.01	1	0	1	0	1	3	1.11
SIT BEHIND	1.64	5	5	2	2	1	4	0
STAND BEHIND	2.93	1	3	2	15	1	5	7.77
LIE BEHIND	4.29	6	5	2	5	1	0	2.22
SIT OVER	0.85	2	1	1	2	0	0	2.22
STAND OVER	2.53	1	3	1	2	1	3	4.44
LIE OVER	0.73	0	1	1	1	0	1	1.11
SIT ABOVE	0.2	0	0	0	1	0	1	0
STAND ABOVE	0.65	2	0	3	1	0	0	1.11
LIE ABOVE	0.45	3	0	1	0	0	1	0
SIT UNDER	0.82	0	0	0	1	2	3	1.11
STAND UNDER	0.79	1	2	2	0	2	0	1.11
LIE UNDER	1.02	2	2	1	1	1	1	0
SIT UNDERNEATH	0.07	0	0	0	0	0	0	0
STAND UNDERNEATH	0.05	0	0	0	0	0	0	0
LIE UNDERNEATH	0.24	0	0	0	0	1	0	0

Appendix 2. Normalized frequency (per million words) of selected posture expressions in spoken corpora. Verbal lexemes, not just base forms, were used in the searches.

SPOKEN	LLC	COLT	WSC
SIT HERE/THERE	38	178	112
STAND HERE/THERE	4	58	27
LIE HERE/THERE	0	4	6
SIT IN	40	68	70
STAND IN	16	12	15
LIE IN	10	16	13
SIT ON	22	78	58
STAND ON	18	22	16
LIE ON	2	10	13
SIT AT	8	10	15
STAND AT	10	4	11
LIE AT	0	0	0
SIT WITH	4	14	7
STAND WITH	4	6	4
LIE WITH	0	0	2
SIT BY	0	0	2
STAND BY	2	0	7
LIE BY	0	0	0

SIT NEXT TO	2	50	5
STAND NEXT TO	2	2	0
LIE NEXT TO	0	0	0
SIT BESIDE	0	0	1
STAND BESIDE	0	0	3
LIE BESIDE	0	0	0
SIT BEHIND	0	0	6
STAND BEHIND	4	8	2
LIE BEHIND	4	0	2
SIT OVER	2	8	3
STAND OVER	0	4	2
LIE OVER	0	0	0
SIT ABOVE	0	0	0
STAND ABOVE	0	0	0
LIE ABOVE	0	0	0
SIT UNDER	0	0	1
STAND UNDER	0	0	1
LIE UNDER	0	0	1
SIT UNDERNEATH	0	0	0
STAND UNDERNEATH	0	2	0
LIE UNDERNEATH	0	0	0