Unit 13 Lexical and grammatical studies

13.1 Introduction

We noted in units 10.2 and 10.3 that lexical and grammatical studies are probably the areas that have benefited most from corpus data. This unit presents four excerpts which discuss the use of corpora in these areas. The first excerpt, Krishnamurthy (2000) demonstrates the use of corpora in collocation analysis, which is followed by an excerpt from Partington (2004), which studies semantic preference. For grammatical studies, Carter and McCarthy (1999) explore the English *GET*-passive in spoken discourse while Kreyer (2003) compares genitive and *of*-construction in written English.

13.2 Krishnamurthy (2000)

In this excerpt, Krishnamurthy briefly reviews the British tradition of text analysis, from Firth to Sinclair, which is closely tied to collocation. On the basis of this review, Krishnamurthy discusses the relevance of corpus linguistics to collocation analysis. This excerpt provides background knowledge for case study 1, where we will explore collocations in the BNC and discuss the implications of our findings for pedagogical lexicography.

Krishnamurthy, R. 2000. 'Collocation: from *silly ass* to lexical sets' in Heffer C., Sauntson H. and Fox G. (eds.) *Words in Context: A tribute to John Sinclair on his Retirement*. Birmingham: University of Birmingham.

2 Collocation theory

We are generally indebted to Firth for channeling the attention of linguists towards lexis (Halliday 1966: 14) and specifically for originating the concept of collocation. Writing from the perspective of stylistics, and viewing meaning as dispersed in a range of techniques working at a series of levels, Firth said: 'I propose to bring forward as a technical term, meaning by *collocation*, and to apply the test of *collocability*' (1957: 194).

Firth established the distinction between cognitive and semantic approaches to word-meaning on the one hand, and the linguistic feature of collocation on the other (196): 'Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation with *night*.' He then proceeded to develop the notion of collocation with reference to examples from specific registers, genres, authors, and texts.

His first example concerned the meaning by collocation of *ass* in the colloquial English of his day, in sentences like 'You silly ass!'. He suggested that the set of potential adjectives with *ass* was limited (e.g. *silly, obstinate, stupid, awful;* and *young* rather than *old*) and that the plural form *asses* was not very common in this meaning.

In the next example, he discussed collocation in the language of Edward Lear's limericks and noted that '*man* is generally preceded by *old*, never by *young*…One of the "meanings" of *man* in this language is to be immediately preceded by *old* in collocations of the type, *There was an Old Man of…* The collocability of *lady* is most frequently with *young*, but *person* with either *old* or *young*. In this amusing language, there is no *boy* or *young man* or *woman*, neither are there any plurals for *man, person,* or *lady*.'

Throughout his discussion of collocation, Firth implied a quantitative basis for the notion, stating actual numbers of occurrences for words in Lear's limericks as well as using

expressions like *habitual, commonest, frequently, not very common, general, usual* and *more restricted*.

Halliday identified the need to measure the distance between two collocating items in a text: 'some measure of significant proximity, either a scale or at least a cut-off point' (1966: 152). Importantly, he brought in the concept of probability, thereby validating the need for data, quantitative analyses, and the use of statistics: 'The occurrence of an item in a collocational environment can only be discussed in terms of probability.' (159) He also first suggested the idea of using collocation to identify lexical sets, which will be explored further at the end of this paper: 'It is the similarity of their collocational restriction which enables us to consider grouping lexical items into lexical sets.' (156)

Sinclair had already started to collect data and perform quantitative analyses using computers in the 1960s. He devised computational methods of looking at collocation in a corpus, and introduced the parameter of position (Sinclair *et al* 1970: 8): 'Collocations of very frequent words are positionally restricted… Collocation which is positionally free… will commonly be an indication of lexical patterning' [Sinclair *et al* 1970 is now reprinted as Krishnamurthy (ed.) 2004]. The very frequent words are of course mainly grammatical words, and Firth had already suggested the separate term *colligation* for collocations involving these. It is the positionally free i.e. lexical, co-occurrences which are now usually termed *collocations*.

In later work, Sinclair presented the 'open choice principle' and the 'idiom principle' as two simultaneously available speaker strategies, with collocation as an important aspect of the latter (1987b: 325): 'Collocation…illustrates the idiom principle. On some occasions, words appear to be chosen in pairs or groups and these are not necessarily adjacent.' He used the corpus frequency of node and collocate to distinguish between *downward* collocation, involving a more frequent node A with a less frequent collocate B, and *upward* collocation: 'Upward collocation of course is the weaker pattern in statistical terms, and the words tend to be elements of grammatical frames, or superordinates. Downward collocation by contrast gives us a semantic analysis of a word.' (326)

3 Corpus linguistics

'In the past, linguists and lexicographers relied mostly on native-speaker intuitions. These were often incorrect, or at least inexact, because each of us has only a partial knowledge of the language, we have prejudices and preferences, our memory is weak, our imagination is powerful (so we can conceive of possible contexts for the most implausible utterances), and we tend to notice unusual words or structures but often overlook ordinary ones' (Butterfield and Krishnamurthy forthcoming) [later published as Butterfield and Krishnamurthy (2000)]. To this one can add that intuition-based made-up examples in dictionaries have been shown to be poor, especially as regards collocation (Krishnamurthy 1996: 145-6; 1997: 45-6, 54-5).

Technological advances in computers, the assembling of large language corpora, and the development of computational techniques for corpus analysis have made possible performance-oriented descriptions of language based on data, in contrast to the previous competence-oriented, intuition-based ones. Indeed, as English becomes a global language, no individual or group can keep up with language change on a worldwide scale; and with the arrival of the Internet in particular, these changes are taking place too fast for our intuitions to assimilate them (Butterfield and Krishnamurthy forthcoming).

Hunston and Francis (2000: 14) equate Sinclair's work with corpus linguistics and argue that it 'prioritises a method, or group of methods, and a kind of data rather than a theory.' The central position occupied by 'a method, or group of methods' in corpus linguistics suggests that advances in methodology will represent progress in this field.

The first step in the established methodology is pattern recognition, to identify the objects of study. For example, the corpus linguistics definition of a word is usually 'a sequence of characters bounded by spaces.' Though this definition may be unsatisfactory, it allows us to proceed in a reasonably objective manner. Other classificatory tasks have been carried out successfully, such as the morphological grouping of word-forms into lemmas (lemmatisation) and the syntactic grouping of word-forms into word-classes (part-of-speech tagging), and

semantic and pragmatic annotations of corpora are currently being attempted, but the fundamental weakness of these operations is that they crucially depend on pre-existing ideas about language.

The second step in the methodology is the generation of frequency lists, which enable us to place the objects of study in a hierarchy. If we accept that the rank of a word-form in a corpus frequency list has some relationship to the importance of that word-form in the linguistic system, frequency lists enable us to decide, on a more objective basis, which word-forms to analyse.

Concordances represent the third stage in the methodology. Concordances allow us to observe the behaviour of a particular word-form in detail. The close inspection of numerous examples of a word-form, together with some context from the original source text, will probably always remain the ultimate process of verification for attributing any linguistic feature to that word-form. As regards collocation, the ability of the computer to sort the examples alphabetically by the words occurring to the left or right of the target word (keyword or node) enable us to notice more easily patterns involving adjacent words, but collocates at a greater distance can still be difficult to spot. And it is impossible for the human brain to distinguish 'statistically significant' collocations merely from the patterns observed in the concordances.

Collocational tools offer the fourth objective methodological element in corpus analysis. Collocation is among the linguistic concepts which have benefited most from the advances in corpus linguistics. Only with large-scale computer corpora can we raise the status of collocation beyond the simple definition 'the co-occurrence of two lexical items in a text within a certain proximity,' which privileges every vagary of performance-error, idiolect or creativity equally, and establish collocation as a powerful organisational principle of language.

With large corpora, we can discuss 'statistically significant' collocations, both those which operate across all language types, as well as those which are restricted in their distribution to particular linguistic modes, genres, varieties, and so on. Computer-generated lists of collocates for a word-form, whether based on raw frequency or more sophisticated statistical measures, focus our attention on prominent candidates for significant collocation, and collocation profiles including positional information make more detailed analysis of syntagmatic or phraseological units possible.

Using the ideas expressed in the academic literature described in Section 2, we might state the primary requirements of collocational software as follows: a) co-occurrences should be identifiable; b) the span or window of collocation (i.e. the distance between collocating items) should be specifiable and adjustable; c) the frequency of co-occurrence should be calculable; d) statistical measures that enhance the notion of 'significance' should be available; e) positional information i.e. where the collocate occurs with respect to the node, should be available; f) if one wants to distinguish between colligation and collocation, it should be possible to distinguish between grammar words and lexical words among the collocates; g) it should be possible to check the distribution of a collocate across source texts to see whether it is restricted to a particular author, mode, vintage, genre, variety, domain etc.

13.3 Partington (2004)

Semantic prosody and semantic preference are two important concepts related to collocation (see unit 10.2). Semantic prosody refers to the collocational meaning hidden between words, or in Louw's (2000: 57) terms, 'a form of meaning which is established through the proximity of a consistent series of collocates.' For example, *HAPPEN* typically refers to some unpleasant situation. Semantic preference, on the other hand, refers to the frequent co-occurrence of a lexical item with a group of semantically related words, or so-called 'lexical set' (Hoey 1997: 3). In this excerpt Partington explores semantic preference in English.

Partington, A. 2004. '"Utterly content in each other's company": Semantic prosody and semantic preference'. *International Journal of Corpus Linguistics* 9/1: 131-156.

2. Semantic preference

2.1 Introduction and definition

Stubbs (2001) following Sinclair (1996, 1998) lists four separate kinds of relation between lexical units, in ascending order of abstraction:

(i) collocation: the relationship between lexical item and other lexical items;

(ii) colligation: the relationship between lexical item and a grammatical category: "For example, the word-form *cases* frequently co-occurs with the grammatical category of quantifier, in phrases such as *in some cases, in many cases*" (Stubbs 2001:65);

(iii) semantic preference;

(iv) semantic (Stubbs prefers the term *discourse*) prosody.

The third of these, *semantic preference*, is defined by Stubbs as "the relation, not between individual words, but between a lemma or word-form and a set of semantically related words" (2001:65). He cites the item *large* which often co-occurs with words for "quantities and sizes", such as *number(s), scale, part, amounts, quantities*. Elsewhere he explains that an item shows semantic preference when it co-occurs with "a class of words which share some semantic feature (such as words to do with 'medicine' or 'change')" (2001:88).

Partington (1998:34–39) examines the intricate semantic preferences of the item *sheer*, an intensifying adjective. This study also illustrates how the typical syntactic realisations of the phrases involving this item are interdependent on these meaning preferences. A concordance of the item was prepared from the newspaper and academic corpora combined (see the beginning of Section 1.2 above). Further analysis showed how it collocated with a number of items from specific semantic sets. These included (i) "magnitude", "weight" or "volume", e.g. *the sheer volume of reliable information, and sheer size of the stadium*, (ii) items expressing "force", "strength" or "energy", e.g. *the sheer force of his presence*. A typical phraseology here was found to be "*the sheer* (magnitude or force word) *of* (noun phrase)".

A third group consisted of words expressing (iii) "persistence", e.g. "*sometimes through sheer insistence*". But here the typical structure is not "*the sheer* (noun phrase) *of* (noun phrase)". Instead we find *sheer* frequently preceded by words expressing means or manner, e.g. *through, out of, by, because of, by virtue of*.

In yet another group, *sheer* collocates with nouns expressing (iv) "strong emotion", e.g. *sheer joy in life, in moments of sheer exhilaration* and (v) physical quality e.g. *he didn't have* […] *the sheer glamour of evil*. This use is often found as part of a list of qualities or emotions, e.g. *nothing can replace the skill, wit, or sheer expertise* […].

Partington (1998:39–47) then goes on to compare this behaviour with other items sometimes considered synonymous with *sheer*, such as *complete, pure* and *absolute*, and discovers that none of them share these semantic preferences and are involved in different typical syntactic structures.

2.2 Group preference

In the previous section we explored the syntactic-semantic conduct of a single item. In another, earlier study using the then 10 million-word *Cobuild* corpus of general English, Partington (1991) examines the collocational behaviour of the group of items called *maximizers* by Quirk et al. (1985), a subset of *amplifying intensifiers*. They include *absolutely, perfectly, entirely, completely, thoroughly, totally* and *utterly*. The first of these, *absolutely*, displays a distinct semantic preference in collocating with items which have a strong or superlative sense: among its significant collocates (i.e. those which co-occur with the keyword three times or more) in the *Cobuild* corpus were: *delighted, enchanting, splendid, preposterous, appalling, intolerable*. There appears to be an even balance between favourable and unfavourable items. This preference is well documented in modern corpus-based dictionaries: "*Absolutely* can be used to add force to a strong adjective" (*Cambridge International Dictionary of English* 1995:5).

Perfectly, on the other hand, exhibits a distinct tendency to co-occur with "good things", including:
capable, correct, fit, good, happy, harmless, healthy, lovely, marvellous, natural
but the fact that we also find *ridiculous* and *odd* demonstrates that language users are able to swim against the current – can "switch off" primings – when they seek particular creative effects. Semantic prosody and preference do not ordain that counter examples *cannot* happen, just that they *seldom* happen. And the value and status of such counter instances could not be more momentous. As Hoey (forthcoming) [later published as Hoey (2004)] tells us "fluency comes from conformity to them" whereas "creativity comes from the switching off of primings".

Of particular interest to the present study, however, is the behaviour of a subgroup within the maximizers which consists of *completely, entirely, totally* and *utterly*. These four have a great deal in common, in particular, they share a large number of collocates (i.e. there is a high degree of *collocational overlap* amongst them). In contrast there is very little collocational overlap between these items and *absolutely* or *perfectly* or, indeed, any of the other items listed in Quirk et al as amplifiers.

A closer investigation of their individual significant collocates was most intriguing. If we examine *utterly*, which is, as it were, the "purest specimen" of the group, we discover that the modified items almost invariably express either the general sense of "absence of a quality" or some kind of "change of state". Those collocates which fall into the first category are: *helpless, useless, unable, forgotten*; those in the second category are: *changed, different*; whereas *failed, ruined, destroyed* seem to fall into both categories. Only two collocates could be interpreted as having a semantic element of "favourable" – *pleasant* and *clear*. This data coincides with Greenbaum's (1970:73) and Louw's (1993:160–161) observation that *utterly* tends to have unfavourable implications. This said, the existence of *utterly pleasant* demonstrates again how tendencies are not inviolable and can be exploited by speakers for particular effects, e.g. (from the *Cobuild* written corpus):
(58) Their relationship in fact was so complete that they were *utterly content in each other's company*.

However, the semantic preference of *utterly* for "absence" and "change" is more fundamental still. Of course, these preferences and the bad prosody may well be connected. In universal terms, in human psychology the presence of something is preferable to its absence (to have is better than to have not). In equally universal terms, change causes anxiety to the human psyche (see Hunston (forthcoming) [later published as Hunston (2004)] on the value of *change*).

Turning to *totally* we find again a large number of "absence" or "lack of" collocates:
bald, exempt, incapable, irrelevant, lost, oblivious, uneducated, unemployed, unexpected, unknown, unpredictable, unsuited, ignored, excluded, unfamiliar, blind, ignorant, meaningless, unaware, unable, vanished, naked, without
and also a set of "change of state", "transformation" words:
destroyed, different, transformed, absorbed, failed

Completely co-occurs with the following "absence" words:
devoid, disappeared, empty, forgotten, hopeless, ignored, lost, oblivious, vanished, gone, lacking, unexpected, bald, naked, unaware, innocent, blank, unknown, finished (and, perhaps, *dry, alone*)
and these expressing "change":
altered, changed, destroyed, different.

For *entirely* we find, in the first category:
different, forgotten, unrelated, without, abandoned, unfamiliar, lost, ignorant, lacking, unnecessary, unknown (and, perhaps, *alone, isolated*)
for "change" we have *different*.

The collocates of *entirely*, however, also seem to encompass a slightly wider range of senses than those of the others. They include a number of words which express an opposition between dependence-independence or relatedness-unrelatedness:

dependent, due, self-sufficient, unrelated, without, isolated.

One last maximizer to deserve a mention is *thoroughly*. This was found in the company of words relating to emotions and states of mind: *annoyed, approved, enjoyed, confused, happy, sure, disgruntled*. It also, strikingly, often co-occurred with words which have something to do with water and washing: *wet, dry, absorbed, cleaned, filtered, muddied* as well as the verbs *wash* and *rinse*. This amplifier evidently retains traces of its ancient sense of *thorough-like*, of penetration, and both water and emotions penetrate "through and through".

We might summarize these observations as follows:

Maximizer:	Preference for:	Prosody:
absolutely	hyperbole, superlatives	
perfectly		favourable
utterly	absence/change of state	unfavourable
totally	absence/change of state	
completely	absence/change of state	
entirely	absence/change of state, (in)dependency	
thoroughly	emotions/liquid penetration	

So far, then, we have found evidence of the existence of the following major types of semantic preference: "factual" – "non-factual", "absence", "change", "emotions" and perhaps "dependence" – "independence". This suggests that there may well be closer links than is generally assumed between grammar and lexico-semantic features, in the sense that the two areas may share a number of similar categories and dichotomies. "Absence", as a marked lexico-semantic feature contrasting with the unmarked feature "presence", would seem reminiscent of negativity, marked in relation to the unmarked positive, in the grammatical system of *polarity*. In addition, the averral of factuality or non-factuality is also accomplished grammatically through choices in the system of verbal epistemic *modality* (with the major difference, however, that the latter is scalar rather than polar or antithetical). Finally, dependence and independence are the two basic relations between clauses. We may tentatively surmise that more particular clausal relations such as causality, contrast and hypotheticality also have their expression at the level of semantic preference.

We hypothesize, then, that a number of the major grammatical functions are reflected at the lexical level in the phenomenon of preference. This should come as no surprise if we accept that linguistic categories are largely modelled by functional communicative needs. Clearly, the same communicative pressures are active on language at both "higher" grammatical and "lower" lexical levels. One of the great challenges currently facing modern linguistics is to describe the precise relationship between these two levels, and semantic preference may well prove a fruitful area of research.

13.4 Carter and McCarthy (1999)

Carter and McCarthy (1999) discuss the interpersonal meaning of the *GET*-passive in a 1.5-million-word sample from the CANCODE spoken English corpus and outline some implications for an interpersonal grammar. This excerpt is taken from the discussion and conclusion sections of their paper.

Carter and McCarthy 1999. 'The English *get*-passive in spoken discourse: description and implication for an interpersonal grammar'. *English language and literature* 3/1: 41-58.

5 Discussion

The key to understanding the *get*-passive is that it highlights the stance of the speaker in context towards the event and the grammatical subject. It is a clear case where examining sentences and jettisoning the people who produce them and their contexts of production is

inadequate. The *get*-passive might indeed be a linguistic puzzle, but it is considerably demystified the moment we look upon it as something the speaker overlays onto events to mark his/her stance towards those events and their subjects. Some linguists have recognized this, most notably Lakoff (1971), Stein (1979), and Hübler (1991), but the benefit of examining real spoken data is that intuitions on that score can be supported by figures showing actual usage. Our conclusion is that the *get*-passive coincides mostly with adverse or problematic circumstances, but these are adverse/problematic as judged by the speaker. *Get* also coincides overwhelmingly with the absence of an explicit agent, suggesting that emphasis is on the event/process and the person or thing experiencing the process encoded in the verb phrase, rather than its cause or agent. We have also made tentative statements, supported by (albeit limited) statistics, about the frequency of particular verbs, the collocational tendencies these suggest, and the relative absence of adverbials. But in all cases, such statements are no more than probabilities.

At this point we may usefully distinguish between deterministic grammar and probabilistic grammar. Deterministic grammar deals with matters of structural prescription (e.g. that *be*- and *get*-passives are always formed with the past participle of verbs, rather than the base-form or *ing*-form). Such determinism enables grammars of languages to be codified in a relatively straightforward way, and has served linguists well for centuries. Probabilistic grammar is concerned with statements of what forms are most likely to occur in particular contexts of use, and the probabilities may be stronger or weaker. Itkonen's (1980: 338) contrast between 'correct sentences' and 'factually uttered sentences' is apt here. Probabilistic grammars need real corpus data to substantiate their claims, but statistical data alone are insufficient; evaluation and interpretation are still necessary to gauge the form-function relationships in individual contexts, from which probabilistic statements can then be derived. Probabilistic grammar proposals are not new: Halliday (1961: 259) talked of the fundamental nature of language as probabilistic and not as 'always this and never that'. More recently, Halliday has returned to this theme with considerable quantitative evidence from the study of corpora. He is primarily concerned with how frequently the terms in binary grammatical systems (e.g. *present* versus *nonpresent*) actually occur in relation to each other, and argues that the statistical facts of occurrence are 'an essential property of the system – as essential as the terms of the opposition itself' (1991: 31). Halliday recognizes that a probabilistic statement such as 'agentless *get*-passives are nine times more frequent than *get*-passives with agents' has little predictive power, but argues that it is important for interpretation of the choice of form. Halliday (1992) stresses further that the different probabilities of occurrence in different registers is also important, since it is unlikely that terms in opposition will be equiprobable in a corpus reflecting any given register. This is true of a general language corpus as well, and the one must be measured against the other. Several of Halliday's followers within systemic-functional linguistics have also further pursued the matter of unequal probabilities of occurrence of particular forms: Nesbitt and Plum (1988) take a predominantly quantitative line in their study of the distribution of clause complexes in real data, and are interested in what is more likely and less likely to occur, rather than what is possible. Here, though, we go beyond the statistics of occurrence and are interested in the relationship between the probabilism of forms and their relationships to contexts. In the case of our type (a) *get*-passives, the probabilities are that *get* will occur in informal contexts when speakers are marking attitude. Most probably that attitude denotes concern, problematicness in some way, or, at the very least, noteworthiness of the event beyond its simple fact of occurring. Indeed, no deterministic statement about when speakers will choose *get* instead of *be* can be made; judgements about adversativeness, problematicness, noteworthiness, etc. are socioculturally founded and are emergent in the interaction rather than immanent in the semantics of verb choice, or of selection of voice or aspect. Over and above these concerns are the broader questions of probability of occurrence within particular generic types of spoken text: most of our *get*-passives occur in narrating and reporting contexts. It may be possible to show that probabilities of form-functional correlations are an integral characteristic of genres and one of the means whereby they can be adequately described by linguists. Genre derives from

patterned social activity, and those patterns of activity emerge from the accumulated interactions of participants; their linguistic traces are the probabilities upon which grammars can be constructed. Since genres reflect complex activities, the plotting of probabilities should not be mono-dimensional but should investigate the likelihood of different features from different grammatical ranks and categories clustering in any specified context, rather in the way that Biber (1995) demonstrates.

Any grammar which attempts to explicate interpersonal meaning cannot but be probabilistic (i.e. interpretive rather than predictive, to utilize Halliday's terms), since interpersonal meanings are emergent in interaction. This brings us squarely back to our other types of pseudo-passives, (b)-(f) in table 1. The passive gradient itself cannot be prescribed deterministically or predicted absolutely. Possible choices of precise structural configuration as represented in table 1 depend on how the speaker cares to position the subject, event and (possible) agents and circumstances relative to judgements about perceived responsibility and involvement of the participants, the inclusion of essential information, and affective factors such as distaste, humour, amazement, etc. reflecting the speaker's reaction to the events. Within particular genres (e.g. gossip, anecdotes), such affective positioning may be more regularly socioculturally conditioned, and manifested in a greater frequency of occurrence of particular forms.

We can now return to some of our other types of structures and see how they occur in ways that highlight their interpersonal meanings just as we did in the case of the type (a) *get*-passives. (28) shows three different choices of perspective on the verb *frame*, concluding with a type (g) structure:

(28) [Speakers are discussing some photographs]

S1 I'm afraid I can't afford to frame them, but erm …

S2 But do you want them framed?

S1 I'd love to have them framed.

S2 Well if that's the case then the next time we come [S1 Yeah] we'll take them with us [S1 Mm] and then we'll have them framed. [90113001]

In S1's first turn, the simple active is chosen, and agency is ambiguous, though likely to mean 'I cannot afford to pay someone else to frame them', which would be a challenge to S1's positive face (self-esteem) in Brown and Levinson's (1987) terms. S2's response equally avoids explicit mention of agency, thus preserving face (consider the possible alternatives: *Do you want to have them framed? Do you want them to be framed?*, both of which do or could carry greater implications of outside agency), and focuses on the subject and her needs. S1 then openly admits a desire to have an outside agency perform the task, and S2 agrees. Interpersonal equilibrium is maintained, face is preserved, by strategic choices of perspective upon patient and agent.

(29) includes structures of types (b) and (c):

(29) S1 Do you think school had an impact on you?

S2 Massive, massive, erm but I left it all to the last minute. I, you know, I kept telling myself well I'll work in the end, and in the end I did, but it was in the end, very much in the last two, in the last, like in the last couple of days, project work and whatnot. I'd stay up forty eight hours to **get it done** and stuff like that, which er … and I realized, you know, then that, you know, if you put, if I put my mind to what I could do you know … but I realized also that it's much easier if you work all the way through, and I had to **get myself organized** then. There was no point you know just leaving it. I had to do it you know. [90175001]

The speaker is centring himself as the topic, and his actions and their consequences for him. Simple active-voice choices on the two highlighted verb-phrases would have been possible, but would not carry the same affective focus on self-discipline, organization and achievement of the speaker's goals; *get*, with its focus on the subject-as-recipient, expresses these aspects of the speaker's narrative evaluation much more powerfully than simple, canonical active-voice alternatives could have done.

6 Conclusion

An interpersonal grammar, if such is needed (and we would argue that our corpus evidence shows that conventional types of description are inadequate to the task of explicating the difference between the various alternatives on the passive gradient), must necessarily be stated in probabilistic/interpretive terms. This does not weaken such a grammar; on the contrary, it lends strength to the enterprise of examining grammar in context, which many grammarians, especially those working within the field of discourse grammar, are currently engaged in, and offers the possibility of harnessing the power of computerized corpora in the service of qualitative research that reaches beyond the bare statistics of occurrence.

Get-passives and related structures are not the only grammatical features to display strong interpersonal meanings that are best explicated in probabilistic and interpretive terms. McCarthy and Carter (1997) account for right-dislocated elements in this way, using spoken corpus evidence, and McCarthy (1998) investigates a number of grammatical features including speech reporting, tense and aspect, and idiom selection from a similar perspective. We are encouraged in our claims for the *get*-passive by the fact that linguists who have previously investigated the phenomenon have instinctively homed in on features connected with the affective and interactive domains, finding it generally impossible to explain the choice of *get* solely by conventional semantic or syntactic criteria. The present paper has attempted to put the weight of corpus evidence behind other linguists' sound intuitions and to state more precisely the contextual conditions in which the *get*-passive and related forms are likely to occur. Our investigation has, we hope, sharpened the description of the structures and taken a step forward in the understanding of how an interpersonal grammar of English might be formulated.

13.5 Kreyer (2003)

Kreyer (2003) explores the use of genitive and *of*-construction on the basis of 519 instances of the *of*-construction and 179 instances of the genitive in a 45,000-word subcorpus taken from the written section of the BNC. It is argued that the variation of the genitive and *of*-construction can be explained with regard to two major underlying factors, namely 'processability' and the 'degree of human involvement'. This excerpt is from the discussion section of the paper.

Kreyer, R. 2003. 'Genitive and *of*-construction in modern written English: Processability and human involvement'. *International Journal of Corpus Linguistics* 8/2: 169-207.

5. Discussion: processability and human involvement

In this study the variation of genitive and *of*-construction has been analysed with regard to three underlying factors, namely the lexical class of the modifier, the semantic relationship expressed by the constructions, and weight and syntactic complexity. It was an important objective to develop a descriptive framework for each of these factors which, while being sufficient from the view-point of linguistic description, would at the same time allow an efficient statistical analysis. With regard to the first conditioning factor a modified version of Quirk et al.'s gender scale, the personality scale, seemed most suitable. To capture the variety of semantic relationships expressed by the two constructions, a paraphrasation system based on traditional categories was applied. The influence of weight and syntactic complexity was analysed separately for instances of premodification and postmodification. The former could conveniently be described in terms of the number of premodifying items of head and modifier. With the latter, the most common types of postmodification (finite clause, non-finite clause and prepositional phrase) were distinguished from apposition and coordination.

The statistical analysis of the conditioning factors has shown that all of them influence the choice of construction. With regard to the lexical class of the modifier, the analysis essentially supported the ranking of the modifiers as depicted on the personality scale. However, the surprising results for collective modifiers indicated that these, as far as the choice between

genitive and *of*-construction is concerned, tend to be regarded as a non-personal collectivity. The influence of the semantic relationship was demonstrated by the fact that six out of nine categories had a decisive influence on the choice of construction. With many of these categories a preference for certain classes of modifiers was observed, which contributed to the immediate effect of the semantic relationship. The impact of premodification was shown to be most decisive with extremely modifier- and head-heavy constructions, which led to a favouring of the *of*-construction ('mod+3..', 'mod+2') or the genitive ('head+3..'), respectively. With regard to postmodification the data showed that a *proximity-principle* is at work, i.e. those constructions are usually favoured which guarantee that related constituents are in the vicinity of one another.

So far the conditioning factors have, for the most part, been analysed in isolation from each other. Some concluding remarks on the relative power of the different factors and the way they interact might therefore be helpful to draw a more comprehensive picture of the variation of genitive and *of*-construction in modern written English. In Table 9, factor levels with decisive influence on the choice of construction have been ranked according to the preference of *of*-constructions. The line in the middle of the table divides those factor levels which lead to a significant deviance from the corpus norm (genitive vs. *of*-construction: 25.6%/74.4%) in favour of the *of*-construction (upper half) or the genitive (lower half).

Table 9. Hierarchy of decisive factors

Factor level	*of* (Rel. Freq.)
postmodified modifier (N_1)	100.0%
non-personal (lexical)	98.5%
'objective' relationship (semantic)	98.3%
N_1 + 3.. (weight of premodification)	96.8%
'attributive' relationship (semantic)	88.1%
N_1 + 2 (weight of premodification)	84.5%
'partitive' relationship (semantic)	81.3%
personified (lexical)	57.1%
personal common noun (lexical)	49.5%
'origin' relationship (semantic)	40.7%
preposition N_2 (right-branching)	40.5%
'kinship' relationship (semantic)	33.3%
N_2 + 3.. (weight of premodification)	33.3%
'disposal' relationship (semantic)	30.8%
'possessive' relationship (semantic)	23.9%
personal proper name (lexical)	14.9%

The ranking of the factors gives a first impression of their power. Those factors which are at the extreme ends of the hierarchy have the strongest influence on the choice of construction. The pervasive importance of syntactic factors is most obvious with regard to postmodified modifiers, which impose extremely heavy constraints on the choice of construction. These constraints even overrule powerful lexical constraints as can be seen in examples (63) and (64), here repeated as (100) and (101):

(100) The aggressive ardour of the professional golfer who might try to cut the slight dogleg and set himself up for an easier shot into the two-tier green (CS4)

(101) The coming of our friends from the west (ABC)

In these cases, the underlying factor seems to be the requirement of structural and referential clarity, which above has been referred to as the 'proximity principle'. Only if this requirement is met can other factors exert their influence. The lexical class of the modifier shows a decisive influence towards the personal and non-personal poles of the personality scale. With regard to proper names, this factor is only overruled by considerations of weight and syntactic complexity, as the examples above show. Similarly, we find that the occurrence

of genitives with non-personal modifiers is restricted to cases of heavy or complex heads, as in example (34), here repeated as (102):

(102) that term's list of the morning lectures for the first-year undergraduates (AOF)

The influence of weight and syntactic complexity is also strongly felt when semantic relationships that favour the genitive, such as 'kinship' or 'possession', show instances of *of*-construction (examples (103) and (104) ((49), (50)).

(103) the son of the Royal Bucks secretary (CS4)

(104) the realms of more important kings (CB6)

Again, if considerations of weight and complexity are not relevant the usual choice with these kinds of relationships is the genitive.

If we look at the results of this study from a more remote perspective, it seems as if the variation of genitive and *of*-construction can to a large extent be described by two more fundamental principles. The discussion in the previous paragraph has shown that considerations of weight and complexity are decisive with a number of instances and may overrule powerful lexical and semantic factors. This influence can be described by the more general term 'processability'. The second factor, i.e. the semantic and lexical aspect, can largely be described in terms of 'human involvement'. This is obvious with the lexical class of the modifier but also holds true for semantic relationship: mostly we think about 'possession', 'disposal' and 'kinship' in terms of our own species whereas, with partitive or objective relationships, which show a preference for the *of*-construction, the notion of associated human beings is usually very faint. In conclusion, therefore, it might be suggested that a possible descriptive framework for the variation of genitive and *of*-construction could be based on the aspects of 'processability' and 'degree of human involvement'.

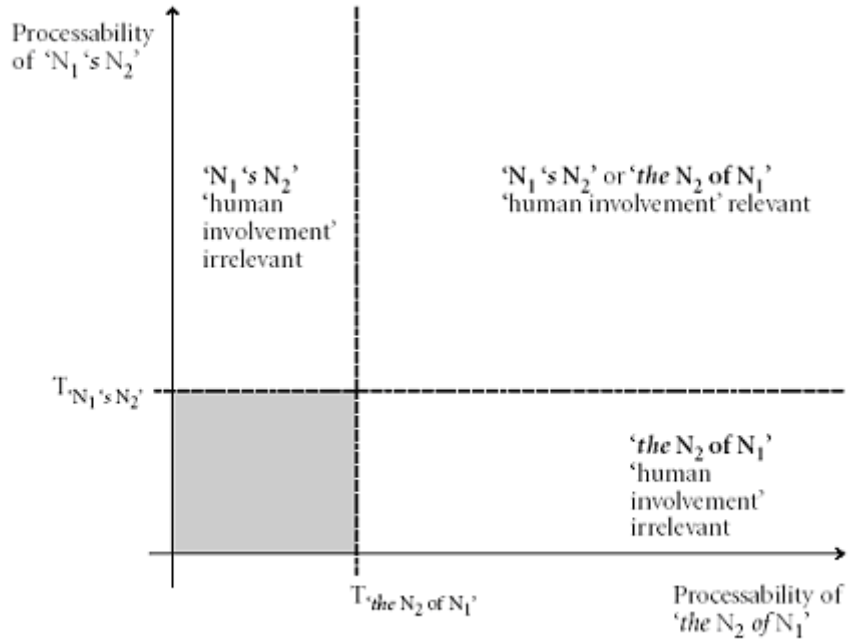

Figure 3. The interaction of 'processability' and 'human involvement'

As to the relative powers of these two factors the data reveal that considerations of 'processability' are more important than the influence of 'degree of human involvement': if, for example, the choice of the genitive led to extreme difficulties of processing, an *of*-construction would be used, regardless of lexical or semantic factors that might indicate genitive. However, instances of heavy modification of head or modifier occur in only 36% of all cases. Although, then, 'human involvement' may not counteract 'processability' in extreme cases, it can exert its influence in almost two thirds of all phenomena, i.e. in those cases where 'processability' is guaranteed by both kinds of constructions ('N_1'*s* N_2' and '*the*

N_2 *of* N_1'). This situation is depicted in Figure 3. The vertical and the horizontal axes show the processability of genitive ('N_1's N_2') and *of*-construction ('the N_2 of N_1'), respectively. For both constructions, there exist threshold levels, $T_{'N1's\ N2'}$ and $T_{'the\ N2\ of\ N1'}$, below which the respective constructions will be extremely difficult to process. In cases where the processability for one construction is below the threshold the alternative construction will be used, regardless of considerations of 'human involvement'; in such cases, this factor is irrelevant. 'Human involvement', however, is relevant in those cases where the choice of construction does not influence processability.

To sum up: in this article, I have attempted to develop a more general description of the variation of genitive and *of*-construction. Starting off from the study of four isolated factors, two basic underlying principles, 'processability' and 'human involvement' have been identified which, to a large extent, account for the variation of genitive and *of*-construction. However, this description can still be further refined since secondary factors, such as information status, might be integrated into the above framework. These factors will most probably show their influence in those cases where 'processability' is guaranteed by both constructions and where 'human involvement' is not decisive. Instances which would be left unaccounted for by the two major factors might then be explained and an even more detailed description of the variation of genitive *of*-construction could be arrived at.

13.6 Unit summary and looking ahead

This unit used four excerpts to illustrate the applications of language corpora in lexical and grammatical studies. In lexical studies collocation and semantic prosody/preference can only be quantified reliably on the basis of corpus data. We will return to lexical studies in case study 1 in Section C, where readers will explore pedagogically oriented corpus-based lexicography. In grammatical studies a corpus-based approach is useful in formulating and testing syntactic theories. It often provides unexpected insights into language use. Readers will have an opportunity to explore, in case study 2 of Section C, the syntactic conditions which may influence a language user's choice between a *to*-infinitive and a bare infinitive following *HELP*. But before that, the next unit discusses the use of corpora in the study of language variation.