Corpus linguistics and English reference grammars

Joybrato Mukherjee

Justus Liebig University, Giessen

Abstract

The present paper begins with a discussion of major conceptual and methodological differences between the new Cambridge Grammar of the English Language *(CamGr), the* Comprehensive Grammar of the English Language *(CGEL), and the* Longman Grammar of Spoken and Written English *(LGSWE). The different approaches in the three grammars are associated with different extents to which corpus data come into play in the grammars at hand. The present paper argues that, for various reasons, the combination of CGEL and LGSWE provides a first important step towards genuinely corpus-based reference grammars in that a theoretically eclectic descriptive apparatus of English grammar is complemented by qualitative and quantitative insights from corpus data. However, there are several areas in which future corpus-based grammars need to be optimised, especially with regard to the transparency of corpus design and corpus analysis and the balance between a language-as-a-whole and a genre-specific description.*

1. Introduction

For a long time, the grammars of the 'Quirk fleet' (cf. Görlach, 2000: 260) have been among the most important reference works in English linguistics. In particular, the *Comprehensive Grammar of the English Language* (CGEL, Quirk *et al.,* 1985) has been widely acknowledged to be the authority on present-day English grammar, bringing together descriptive principles and methods from various traditions and schools in order to cover grammatical phenomena as comprehensively as possible (cf. Esser, 1992). Recent years have seen the publication of two other, similarly voluminous, reference grammars of the English language: the *Longman Grammar of Spoken and Written English* (LGSWE, Biber *et al.,* 1999) and the *Cambridge Grammar of the English Language* (CamGr, Huddleston and Pullum, 2002a). It is both remarkable and telling that both LGSWE and CamGr were mainly inspired by CGEL. In the preface to LGSWE, Biber *et al.* (1999: viii) explicitly refer to CGEL 'as a previous large-scale grammar of English from which we have taken inspiration for a project of similar scope'. As for CamGr, Huddleston and Pullum (2002a: xvi), too, concede that CGEL 'proved an indispensable source of data and ideas'.

Although the genesis both of LGSWE and CamGr is closely linked to CGEL, the descriptions of English syntax that the three grammars offer are fundamentally different from each other. In section 2, I will thus first of all address the question as to what the major conceptual and methodological differences are between the three grammars at hand; in this context, special attention will be paid to the question whether the grammars complement each other or, alternatively, whether they compete with each other. From a corpus-linguistic perspective, it is of course of particular importance to compare the extents to which corpus data are taken into consideration in the grammars under scrutiny. In section 3, I will focus on LGSWE as the first large-scale and fully 'corpus-based' reference grammar and discuss the merits and advantages of this grammar (e.g. its focus on frequencies and its adherence to the descriptive framework set out in CGEL) as well as some areas in which future corpus-based grammars could still be optimised (e.g. with regard to the transparency of corpus design and analysis). In section 4, I will offer some concluding remarks on the usefulness of LGSWE and CGEL as a conjoined reference work for (corpus) linguists.[1]

2. Comparing three reference grammars of English: a reprise

It is of course difficult – if not impossible – to compare in detail the analyses of all grammatical phenomena offered by CGEL, LGSWE and CamGr. However, it is certainly possible and useful to abstract away from the entirety of syntactic analyses the major conceptual, descriptive and methodological differences between the three grammars at hand. Such a comparison was the basis of my review of CamGr (cf. Mukherjee, 2002a), which triggered off a brief – though intense – discussion between the reviewer and the authors of CamGr about all three aforementioned reference grammars.[2] From this discussion, the authors of CamGr themselves derived 'some points of agreement' (Huddleston and Pullum, 2002c). Table 1 provides a somewhat simplistic overview of these points of agreement on general differences between the approaches to English grammar pursued by CamGr, CGEL and LGSWE. To these differences I will briefly turn in the following.

The object of inquiry of CamGr is defined as 'international standard English' (cf. Huddleston and Pullum, 2002a: 4f.). Strictly speaking, then, CamGr is intended to provide the grammar of a specific variety of English (which is used internationally and considered as world standard English). On the other hand, the object of inquiry of CGEL is the so-called 'common core', which 'is present in all the varieties so that, however esoteric a variety may be, it has running through it a set of grammatical and other characteristics that are present in all the others' (Quirk *et al.*, 1985: 16). As pointed out by J. Aarts (2000), however, it is not at all easy to pinpoint exactly this abstract idea of the common core:[3]

> The notion of the common core is an attractive one, but very difficult to operationalize. [...] It is clear that the identification of the common core requires an exhaustive knowledge of all varieties and the ability to tell which of their features they share and which are variety-dependent. For the time being therefore, the notion of a common core must remain an intuitive notion. (J. Aarts, 2000: 19f.)

With the publication of LGSWE, some aspects of the notion of common core are now empirically accessible, because its objects of inquiry are 'four core registers':

Table 1: Some major differences between CamGr, CGEL and LGSWE

	CamGr (Huddleston and Pullum, 2002a)	**CGEL (Quirk *et al.*, 1985)**	**LGSWE (Biber *et al.*, 1999)**
a) object of inquiry	'international standard English'	'common core'	'four core registers'
b) generative influence in general	+	–	
c) preference for binary branching in particular	+	–	
d) preference for multiple analysis and gradience	–	+	–
e) database	intuitive, collected, corpus	intuitive, collected, corpus	LSWE corpus
f) in-depth quantitative analyses	– *	– **	+
	* some corpus-based dictionaries and grammars (and, very occasionally, corpora and archives) were consulted ** some quantitative data from SEU, Brown and LOB were taken into consideration		

'conversation', 'fiction', 'newspaper language' and 'academic prose' (cf. Biber *et al.,* 1999: 24ff.). Despite the obvious problems involved in this register distinction, the objects of inquiry of CGEL (i.e. the variety-independent common core) and of LGSWE (i.e. the variety-dependent features of the four core registers) obviously complement each other.

As indicated in Table 1, generative grammar has exerted an enormous influence on CamGr. As Huddleston and Pullum (2002c) point out, they 'have drawn many insights from generativist work of the last fifty years'. An overt example of this generative influence is its strong preference for phrase structure analyses in general and binary branching in particular. In fact, there are only very few fields in which CamGr deviates from binary branching, the two most important exceptions being coordination (cf. Huddleston and Pullum, 2002a: 1279) and ditransitive verb complementation (cf. Huddleston and Pullum, 2002a: 1038). While CamGr may be regarded as a generatively-oriented reference grammar, CGEL has been labelled most appropriately by Standop (2000: 248) as ‘strukturalistisch-eklektisch’ – i.e. as a grammar that follows the tradition of descriptive structuralist grammars and combines it undogmatically and eclectically with concepts from other linguistic schools of thought.[4] In principle, this also holds true for LGSWE, because it takes over to a very large extent the descriptive apparatus of CGEL (cf. Biber *et al.,* 1999: viii).

With regard to the extent to which gradience and multiple analyses are allowed for, CamGr is also fundamentally different from CGEL. In CGEL, gradience of grammatical categories and the possibility of multiple analyses play a significant role because grammar is viewed as an inherently ‘indeterminate system’ (cf. Quirk *et al.,* 1985: 90). Thus, sentences with prepositional verbs (such as *look after*), for example, are analysed in two different ways in CGEL, cf. Figure 1. Neither of them is considered incorrect.

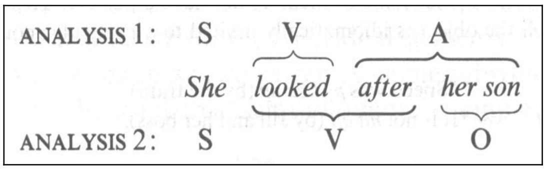


Figure 1: Multiple analysis in CGEL (Quirk *et al.,* 1985: 1156)

CamGr, on the other hand, aims to eradicate as many multiple analyses as possible by positing one specific analysis as correct:

> Quirk et al. tend often to suggest that things are actually indeterminate – vagueness rather than ambiguity, there being no decision about which is the right analysis in some cases. There is an opposite tendency noticeable in The Cambridge Grammar: we try to find arguments that eliminate indeterminacy and home in on a particular analysis, IF the facts can be found to fully support it.
>
> (Huddleston and Pullum, 2002c)

Thus, it does not come as a surprise that Huddleston and Pullum (2002a) forcefully argue that only ‘analysis 1’ in Figure 1 is correct, while ‘analysis 2’ should, in their view, be discarded.[5] It should be mentioned in passing that LGSWE does not place any special emphasis on multiple analyses either, because it usually takes one of the options offered by CGEL as its starting-point for a quantitative analysis.

What clearly emerges from this comparison of some general conceptual and descriptive principles in CGEL and CamGr in particular is the fact that these two grammars are, strictly speaking, not true competitors. Rather, they represent (and put into practice) fundamentally different approaches to English grammar. In other words, it is probably a matter of linguistic ideology (and syntactic taste, if you wish) which of the two grammars one finds more plausible and intuitively appealing. For example, it seems to me that not infrequently (and presumably due to its overall formalist orientation), CamGr succeeds in offering a theoretically amazingly stringent account, but at the expense of breaking with traditional and intuitive analyses. In his review of CamGr, Brdar (2002), for example, refers to the analysis of clauses with auxiliaries:[6]

> [In CamGr] all auxiliaries, primary and modal ones, are effectively treated as main verbs because they are said to take complements in the form of non-finite clauses. This has a number of nasty consequences. First, there are then no complex verb phrases in the sense of exhibiting more than one verb at the same hierarchical level: they either contain a single main verb or an auxiliary plus another verb or verbs as clausal complements at a hierarchically lower level(s). Secondly, an ordinary sentence like:
>
> *He may know her.*
>
> must be analysed as being biclausal, which runs counter to all our intuitions, at least in present-day English.
>
> (Brdar, 2002: 81)

The formalist stringency of the biclausal analysis of clauses with auxiliaries offered by CamGr is innovative and impressive. But the (probably unanswerable) question remains whether the analysis offered by CamGr is inherently better than the traditional analysis suggested by CGEL (according to which ‘may know’ in ‘He may know her’ would be regarded as one complex verb phrase of one clause).

In spite of the clear conceptual and descriptive differences, CamGr and CGEL share one important feature: as shown in Table 1, neither of the two grammars is systematically and entirely based on corpus data and their in-depth quantitative analysis. In CGEL, there are, for example, some references to quantitative data from the Survey of English Usage, e.g. the distribution of noun phrase types across different genres (cf. Quirk et al., 1985: 1350ff.). As Sinclair (1991: 100f.) and Stubbs (1993: 9f.) have already noted, however, CGEL is not systematically – let alone, exhaustively – based on authentic examples from the corpus, but rather on intuitively invented and – possibly – other unspecified (‘collected’) data, including elicitation experiments (cf. Quirk et al., 1985: 33). In CamGr, the authors also base their description on a mix of data, ranging from intuitions and invented data, randomly collected data and elicited data to corpus data:[7]

> The evidence we use comes from several sources: our own intuitions as native speakers of the language; the reactions of other native speakers we consult when we are in doubt; data from computer

> corpora (machine-readable bodies of naturally occurring text), and data presented in dictionaries and other scholarly work on grammar. […] [Apart from computer corpora, we] have also drawn on a variety of other sources, including collections of our own from sources such as magazines, newspapers, plays, books, and film scripts.
>
> (Huddleston and Pullum, 2002a: 11)

In a similar vein to CGEL, the grammatical analysis and theory in CamGr is thus unsystematically related to the corpora that have been used. With regard to corpus data, then, the methodology both of CGEL and CamGr amounts to what Esser (2002: 133) has repeatedly called the 'butterfly method'. On a more negative note, the corpus is used as a quarry, as it were, from which isolated data and frequencies are extracted. Thus, the two grammars may be regarded as 'corpus-aware' but they are, strictly speaking, not 'corpus-based'.

From a methodological perspective, LGSWE is thus clearly different from CGEL and CamGr in that it is entirely and solely based on the 40-million-word Longman Spoken and Written English (LSWE) Corpus, including authentic texts from a wide range of natural communication situations in spoken and written English; all examples are taken from the corpus. It is this approach that Biber *et al.* (1999) call 'corpus-based':

> The LGSWE adopts a corpus-based approach, which means that the grammatical descriptions are based on the patterns of structure and use found in a large collection of spoken and written texts, stored electronically, and searchable by computer.
>
> (Biber *et al.,* 1999: 4)

What makes LGSWE particularly useful and easily accessible is not only the fact that its object of inquiry (i.e. the 'four core registers' of 'conversation', 'fiction', 'newspaper language' and 'academic prose') complements the object of inquiry of CGEL (i.e. the 'common core', see above), but also Biber *et al.*'s (1999: 7) wise decision to take over, with only very few exceptions, the descriptive framework and terminology of CGEL, which without any doubt 'has gained a broad currency through its use in other grammars, textbooks, and academic publications'. Some of the overall advantages of the shared descriptive apparatus and its implications for the future of corpus-based English reference grammars will be taken up in section 4.

Without any doubt, LGSWE is innovative in its focus on corpus data and the corpus-guided analysis of differences between spoken and written language in general and the 'four core registers' in particular. Nevertheless the question arises as to what extent LGSWE represents – or contributes to – a genuinely 'corpus-based' grammar. It is an assessment of this issue – with some suggestions for future optimisation – to which I will turn in the following section.

3. Optimising corpus-based grammars: the *Longman Grammar* and beyond

At the PALC Conference in 1999, J. Aarts (2000: 17) made a plea for a 'new generation of corpus-based English grammars'. On the eve of the publication of LGSWE, he demanded that corpus-based grammars of 'language use' satisfy the following four requirements:

> 1. it should allow the description of the full range of genre varieties and the full range of medium varieties, from spontaneous, non-edited language use (usually spoken), to non-spontaneous edited language use (usually written/printed). […]
> 2. it should allow the combination of a quantitative and a qualitative description of the data. […]
> 3. it must establish a relation between phenomena that are external to the language system on the one hand and system-internal phenomena on the other. […]
> 4. it should allow an integrated description of syntactic, lexical and discourse features.
>
> (J. Aarts, 2000: 26f.)

There presumably is general agreement about the relevance of these four requirements in that they pick up on some of the most important insights into the nature of language use that modern corpora have provided. In this context, it is also beyond reasonable doubt that LGSWE is the first large-scale attempt to meet these standards and put them into practice in a comprehensive reference grammar of the English language: (1) LGSWE analyses four core registers of English and compares spoken and written language; (2) it tries to explain quantitative corpus findings by means of a qualitative discussion of the data; (3) to this end, it takes into account the influence of various discourse factors on linguistic choices; (4) it takes into consideration the interdependence between lexis and grammar by identifying the lexical items that tend to co-occur with specific syntactic structures. I thus regard LGSWE as being a significant step on the way towards a new generation of corpus-based grammars, as envisaged by J. Aarts (2000).

However, despite the positive echo that LGSWE has already found in the linguistic community due to its innovative features (cf. e.g. Carkin, 2000; Görlach, 2000; Krug, 2002; Schmid, 2003), there is of course room for improvement in various regards. There are three areas in particular in which future reference grammars should be optimised and to which I would like to turn in the present paper: (1) the transparency of the database and the analysis, (2) the balance between a language-as-a-whole and a genre-specific description, (3) the openness to constant revision and modification.

3.1 Transparency of the database and the analysis

At first blush, it seems to be banal to demand that the database of a 'corpus-based' grammar and the corpus analysis be made transparent to the user of the grammar. Right from its beginning, the transparency of data and analysis has always been at the heart of modern corpus linguistics, since the size and representativeness of the data, the reliability and the replicability of the analyses were the methodological innovations that set it apart from, say, generative approaches to language. Nevertheless, some users' discomfort with LGSWE is caused by the very lack of the transparency of the data and the analyses in this grammar. A major problem is that users of the grammar are never told which specific texts – or which particular passages from which texts – the LSWE Corpus contains in its entirety; the description of the LSWE Corpus only provides information on the kinds of texts, registers and sub-registers that are included in the corpus and their balance and gives a few text examples of various registers (cf. Biber *et al.,* 1999: 24ff.). Biber *et al.*'s (1999: 24) statement that the 'LSWE Corpus is constructed to provide a systematic representation of different registers' may well be true; however, since they do not give the exact sources of all corpus texts, their claim is simply not testable nor are their findings replicable.[8] Given this lack of testability of the representativeness in corpus design, it comes as no surprise that various reviewers of LGSWE are extremely critical of the definition and demarcation of individual genres, the overall balance of genres, the text selection and the sampling techniques. Consider, for example, Schneider's (2001) critical remarks:

> On the one hand, their [Biber *et al.*'s (1999)] register categories conceal a great deal of internal variation by topic, sociolinguistic background, etc. (news from a tabloid may be expected to follow patterns different from those in the *Wall Street Journal*); on the other hand, the composition of the samples is quite uneven. To some extent, this looks suspiciously (and deplorably) like convenience sampling rather than a principled selection strategy.
>
> (Schneider, 2001: 139)

Whether or not Schneider's (2001) criticism is justified or not – the important point is that the issues he raises cannot be properly addressed and discussed because the large majority of texts included in the LSWE Corpus remain unspecified in LGSWE. This shortcoming – together with the problem of too broadly defined registers such as the news register – gives free rein to even harsher and more fundamental criticism of the corpus design. Parker (2003), for example, calls into question the very representativeness of the LSWE Corpus altogether by stating that LGSWE does not provide a grammatical description of the English language, but only of the corpus it is based on:

> If you would be happy to own a book called *The Longman Corpus of Spoken and Written English* – which is what this book should have been called – then you will be willing to forgive its shortcomings.
>
> (Parker, 2003: 97)

As a matter of fact, this kind of criticism could easily have been countered by basing the grammatical description on a corpus that was not compiled and designed specifically – and idiosyncratically – for LGSWE. For future reference grammars, it may indeed be more useful to use already existing megacorpora (say, of the calibre of the British National Corpus) which have already been widely used, whose sources are absolutely transparent, and on whose representativeness some sort of general agreement has already emerged.

Another shortcoming of LGSWE that adds to the impression of a general lack of transparency is the fact that, not infrequently, the quantitative and qualitative analysis of a particular phenomenon is not based on the 40 million words of the LSWE Corpus in its entirety, but on a very small subcorpus.[9] What is more, the design of the subcorpus usually remains largely unspecified. For example, the quantitative analysis of fronting (cf. Biber *et al.,* 1999: 900ff.) turns out to be based on a subcorpus of 200,000 words (i.e. 0.5% of the LSWE corpus), about which only sketchy details are provided in the corresponding endnote: 'Based on a sample of 200,000 words from the LSWE Corpus: 25 texts of 2,000 words each from conversation (BrE only), fiction, news, and academic prose' (Biber *et al.,* 1999: 1136).[10]

Finally (and this may come as a surprise to corpus linguists), I would argue that the grammatical description could become more transparent and more easily accessible to the user of a corpus-based reference grammar if simplified and invented examples were used in addition to authentic corpus examples whenever necessary. This is a suggestion for improvement that has already been put forward by Parker (2003) in his review of LGSWE; he contrasts the clarity of the invented examples of the use of *some* and *any* in CGEL with the opaqueness of the authentic examples of the adverb position in clausal negation in LGSWE:

> *Some* people *never* send *any* Christmas cards.
> **Any* people *never* send *some* Christmas cards. (Quirk et al., 1985, 85)
> "Our investigations indicate that this substance was **not deliberately** administered." (FICT)
> Alexander looked at Wilkie who **deliberately did not** see him. (FICT) (*LGSWE*, p. 175)
> Imagine you are a nonnative speaker trying to infer the concept of negative scope from these two sets of examples. Even without examining any explanatory text surrounding the sample sentences, the point of Quirk et al. (1985) is clear: in a declarative structure, *any* must follow (i.e., occur within the scope of) a negative element. However, the point of the LGSWE sentences is opaque. (For example, the reader might wonder why did appears in the second sentence but

> not in the first. Is it relevant?) The point is that sample sentences used to illustrate the essential properties of a structure are much clearer if irrelevant variables are extracted and, when appropriate, negative examples are provided for contrast. Both of these goals are possible with constructed data, less so with "live" data.
>
> (Parker, 2003: 93f.)

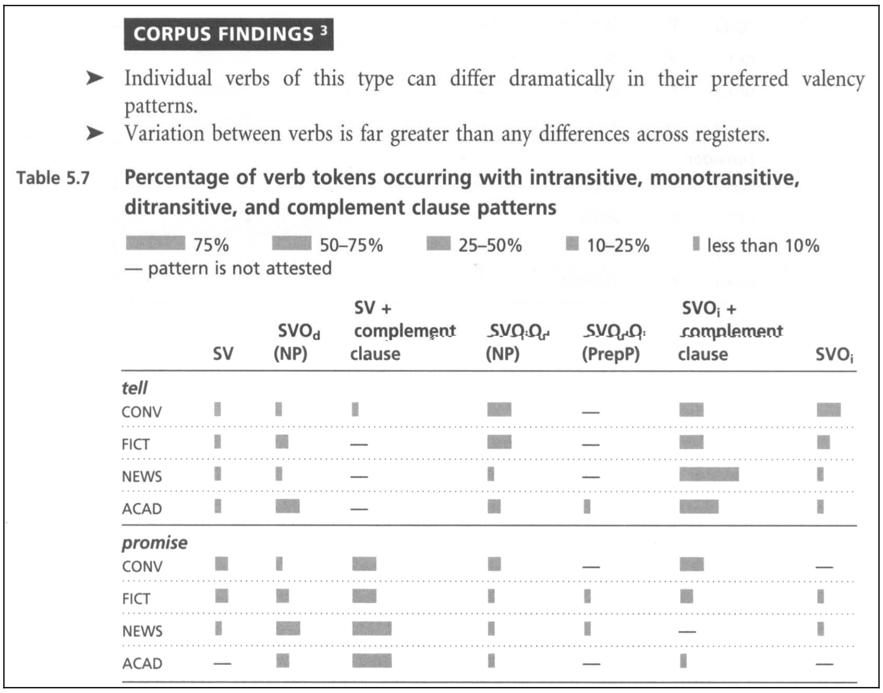
I would also contend that in a reference grammar of the English language – even if it aims to be corpus-based – constructed and/or ungrammatical sentences may fulfil an important function, especially in order to focus on the exemplification of the syntactic phenomenon at hand and in order to define the borderline between what is syntactically possible and what is not. This is not to say that the consistent use of authentic corpus examples is irrelevant; it is not. There are many fields in which authentic corpus examples are much more suitable than artificial examples, e.g. in exemplifying typical lexico-grammatical co-selections, in showing how discourse factors influence linguistic choices and in illustrating natural spoken interaction.

3.2 Balance between a language-as-a-whole and a genre-specific description

Recently, the term 'monolithic grammar' (cf. Conrad, 2000, as quoted by Hunston, 2002: 161, 167) has been used to refer unfavourably to the traditional kind of a general (reference) grammar that does not distinguish between individual genres or registers but attempts to describe the language as a whole. It is thus not surprising that LGSWE, as a corpus-based grammar, tries to overcome the 'monolithic' tradition and is aimed at a consistently medium-specific and genre-sensitive description of English grammar. Note that LGSWE starts off from the strong claim that general grammatical patterns are of clearly less importance than register-specific patterns in English grammar:

> In most cases, it is simply inaccurate or misleading to speak of a general pattern of use for English; instead, each register has distinctive patterns, associated with its particular communicative priorities and circumstances.
>
> (Biber *et al.,* 1999: 24)

However, LGSWE itself provides a multitude of examples of grammatical patternings that are largely independent of register differences, for example in the case of ditransitive verbs and their preferred complementation patterns.[11] Consider Figure 2, which is taken from LGSWE: it reports on the frequency and distribution of the complementation patterns of the ditransitive verbs *tell* and *promise* in the four core major registers of the LSWE Corpus.[12] This is but one example that illustrates that 'variation between verbs is far greater than any differences across registers' (Biber *et al*., 1999: 388).[13]



CORPUS FINDINGS [3]

- ➤ Individual verbs of this type can differ dramatically in their preferred valency patterns.
- ➤ Variation between verbs is far greater than any differences across registers.

Table 5.7 **Percentage of verb tokens occurring with intransitive, monotransitive, ditransitive, and complement clause patterns**

█████ 75% ████ 50–75% ███ 25–50% ██ 10–25% ▌ less than 10%
— pattern is not attested

	SV	SVO$_d$ (NP)	SV + complement clause	SVO$_i$O$_d$ (NP)	SVO$_d$O$_i$ (PrepP)	SVO$_i$ + complement clause	SVO$_i$
tell							
CONV	▌	▌	▌	███	—	███	███
FICT	▌	██	—	███	—	███	██
NEWS	▌	▌	—	▌	—	█████	▌
ACAD	▌	███	—	██	▌	████	▌
promise							
CONV	██	▌	███	██	—	███	—
FICT	██	██	███	▌	▌	██	▌
NEWS	▌	███	████	▌	▌	—	▌
ACAD	—	██	████	▌	—	▌	—

Figure 2: Frequency and distribution of the complementation patterns of tell and promise (Biber et al., 1999: 388)

It seems to me that the days of general, language-as-a-whole (‘monolithic’, if you wish) grammars are not at all over. Rather, general grammars of the common core need to be complemented – not replaced – by genre-specific descriptions (for which corpus data are of course a great boon). The combination of the language-as-a-whole (and non-corpus-based) CGEL and the register-oriented (and corpus-based) LGSWE provides a good example of how the two perspectives on English grammar may complement each other.

3.3 Openness to constant revision and modification

One of the most essential principles in corpus linguistics is to aim at a description that is true to the facts of actual usage. Probably all corpus linguists subscribe to Sinclair’s (1991: 4) point of view that it is necessary for linguists to ‘accept the evidence’ and ‘reflect the evidence’ (with ‘evidence’ meaning actual data), implying that existing and intuition-based descriptions (say, of grammar) have to be modified or even revised if they turn out not to account for actual usage as attested in corpus data. However, it is quite clear that the concept of actual usage is a moving target because language – including grammar – changes continuously (cf. Mair, 2002). In order to keep track of ongoing change in the English language, Sinclair (1991: 26) has propagated the idea of a ‘monitor corpus’ –

resulting in the dynamic Bank of English Corpus. We might thus envisage some sort of corpus-based 'monitor grammar' of the English language that could constantly and speedily be updated, modified and revised if and when changes in English grammar can be traced in new data that are being included in the underlying corpus.[14] For example, the so-called 'double copula construction' as in *the question is is that…* (cf. Andersen, 2002), which has emerged fairly recently and is now increasingly used both in spoken and in written language, could be accounted for by a monitor grammar as soon as the construction is attested frequently enough in the incoming corpus data.

Very often, however, it is not an entirely new grammatical form that has to be included in a corpus-based reference grammar, but it is the grammatical analysis of a well-known structure that needs to be modified in the light of corpus data. In this context, the combination of CGEL and LGSWE already points up some interesting avenues for future work. A good case in point is the extraposition of clausal subjects. CGEL starts off from the traditional distinction between the 'canonical' non-extraposed (and less usual) form and its non-canonical extraposed (and more usual) variant:

> To hear him say that + surprised me ~ It + surprised me + to hear him say that
> But it worth emphasizing that for clausal subjects […] the postponed position is more usual than the canonical position before the verb […].
> (Quirk *et al.,* 1985: 1392)

This account is the intuition-based, pre-corpus point of departure for LGSWE, which takes over the descriptive categories from CGEL. Besides giving more precise frequency information and authentic examples, LGSWE gives various reasons why the extraposed variant is much more frequent in the first place (although the non-extraposed variant is analytically simpler and thus syntactically 'unmarked'), including discourse and processing factors and different production constraints in speech and writing (cf. Biber *et al.,* 1999: 724ff.). The important point here is that the discussion of the corpus findings in LGSWE culminates in a re-categorisation of what is to be considered 'unmarked':

> Extraposed constructions should be regarded as the unmarked choice whenever a *to*-clause functions as logical subject of a main clause.
> (Biber *et al*., 1999: 725)

The description of extraposed subject *to*-clauses as the 'unmarked choice' points to a partial modification of the received account of extraposition because this terminology implies that canonical structure in terms of analytical simplicity on the one hand and the default variant in terms of frequency on the other do not come into operation along with each other in this case: the syntactically simplest structure is clearly outnumbered by a more complex, seemingly derivative variant. Thus, the account in LGSWE is clearly reminiscent of Mair's (1990: 34) view that ‘absence of extraposition can be regarded as a type of “fronting”’. What the combination of CGEL and LGSWE brings to the fore is the fact that there seem to be two different levels of basicness involved in the grammatical analysis of extraposition of clausal subjects and their non-extraposition. Future reference grammars will have to look out for such states of affairs in more detail and try to take into consideration much more systematically the levels of analytical simplicity on the one hand and unmarked choices in terms of frequency on the other.[15]

4. Concluding remarks

In the light of the discussion in the previous sections, the combination of CGEL and LGSWE as complementary grammars may well be regarded as the first significant landmark on the way ‘towards a new generation of corpus-based grammars’, as envisaged by J. Aarts (2000). It is obvious that LGSWE is not a classic stand-alone reference grammar and is heavily dependent on the model and description set out in CGEL. In actual fact, my impression is that most users of LGSWE usually consult this grammar side by side with CGEL, because it is the combined use of CGEL and LGSWE that ensures that two equally important aspects of grammar are covered: (1) the comprehensive – and thus not necessarily and entirely corpus-based – description of the grammatical structures that are possible and the demarcation from those structures that are not admissible in English; (2) the corpus-guided focus on routines (e.g. lexico-grammatical co-selections) and genre-specific trends that are typical of language use. Both aspects of grammar should also be taken into consideration by future corpus-based grammars.

I have tried to sketch out some fields in which future reference grammars may well be improved and could profit from the merits and shortcomings of recent reference grammars of English. By definition, a reference grammar for the widest possible target audience – including linguists and laymen alike – always represents a compromise between in-depth analysis and simplifying generalisation, between formal systematisation and functional interpretation, between objective description and intuitive appeal and between theoretical stringency and easy accessibility. The advent of modern corpora does not solve the problem of how to strike the right balance between these conflicting aims because it adds yet another dimension to this very problem: what is the role of corpus data and of corpus analyses in a corpus-based reference grammar? Generally speaking, I would argue that in contrast to the revolutionary impact of corpora on English linguistics, the development of corpus-based reference grammars of the English language will probably turn out to be much more evolutionary and incremental. On the merely pragmatic side, such an evolutionary process – from CGEL via LGSWE to future work – could well lead to a wider and more long-term acceptance of the use of corpus data in reference grammars than attempts to come up with much more radically new reference grammars.

Notes

[1] With regard to several issues raised in the present paper, I have profited from discussions with various colleagues (not always resulting in a consensus, though). I am particularly grateful to Jan Aarts, Jürgen Esser, Sebastian Hoffmann, Stig Johansson, Geoffrey Leech, and Geoffrey Pullum.

[2] It is not my intention to replicate here all the arguments and counter-arguments put forward in the discussion. For details see – in chronological order – Mukherjee (2002a), Huddleston and Pullum (2002b), Mukherjee (2002b), Huddleston and Pullum (2002c). For a much more detailed comparison of CGEL and CamGr see Leech (2004).

[3] It should be noted that LGSWE was published too late for J. Aarts to include it in his survey of English grammars (cf. J. Aarts, 2000: 17).

[4] In fact, CGEL may well be seen as the culmination of the so-called ‘Great Tradition’ (cf. F. Aarts, 1975: 98).

[5] Note, however, that Huddleston and Pullum (2002a) do not use the term ‘adverbial’ for the post-verbal constituent ‘after her son’ (as in CGEL), but label it a ‘complement’ (as already envisaged by Huddleston (1988) in his critical review of CGEL).

[6] See also Huddleston and Pullum’s (2003: 67) response to Brdar (2002) in which the authors of CamGr explicitly state that ‘[r]igorous analysis of the available syntactic evidence can reveal where our ingrained intuitions about grammar are simply wrong’. This is no doubt true from a strictly syntactic point of view, but the question remains whether a reference grammar of the English language should be based on a ‘rigorous’ analysis, based on formal criteria and unrelated to any kind of intuitively appealing plausibilities.

[7] It should be noted in passing that Huddleston and Pullum (2002a: 11) frequently mention ‘evidence’ as one of the cornerstones of the grammar description offered in CamGr: ‘Issues of interpretation often arise. But always, under the descriptive approach, claims about grammar will depend upon evidence’. Without getting into details, this terminology poses two related problems: first, it does not distinguish between ‘evidence’ and (different kinds of) linguistic ‘data’; secondly, the implicit assumption is that the careful consideration of the ‘evidence’ will result in a specific and correct analysis of a grammatical phenomenon, while all other alternative analyses can be discarded (as falsified by the ‘evidence’). In my view, it would also be sensible for corpus linguists to be more reluctant to use the term ‘evidence’ (and equate it with corpus data).

8 It goes without saying that this critique also applies to other corpus-linguistic resources such as corpus-based dictionaries. It is a pity, in my view, that only very few of them measure up to the exemplary transparency of the corpus design in the first edition of the *Collins COBUILD English Dictionary* (Sinclair, 1987), which explicitly lists all text sources.

9 It is more than unfortunate that in LGSWE the information about whether an individual analysis is based on the full corpus or a subcorpus (and on which subcorpus) is hidden away in the endnote section at the end of the grammar (cf. Biber *et al.,* 1999: 1133ff.). Since LGSWE usually gives relative frequencies (as percentages) or standardised frequencies (per million words), the casual reader may never become aware of the fact that the grammar is not entirely based on an exhaustive analysis of the LSWE corpus.

10 Note in this context Sinclair's (2002: 357) criticism of 'the unexplained selection procedures for what to present and what to leave out, the silence about the huge tidying job that needs to be done to achieve such neat presentation, and the relation of design and comment to implied norms'.

11 I have provided a detailed discussion of the issue of a language-as-a-whole and a genre-specific description of ditransitive verbs elsewhere (cf. Mukherjee, 2005: 112ff.).

12 Note that the corpus findings in Figure 2 are '[b]ased on interactive coding and computer analysis of a random sample of 200 occurrences for each verb from each of the four core registers in the LSWE Corpus' (Biber *et al.,* 1999: 1134). The findings are not replicable for LGSWE users (cf. section 3.1).

13 Another problem of an entirely register-focused/genre-focused approach to grammar is of course that the boundaries that are drawn between registers as well as the definition of individual registers can always be criticised. What is more, a register or genre may also well be seen as a 'monolithic' abstraction in a similar vein to the notion of the English language (cf. Hunston, 2002: 161).

14 In the discussion panel on grammar and corpus linguistics at the 24th ICAME Conference, chaired by Jan Aarts and reported on in this volume, the idea emerged that electronic media could be used for future reference grammars in order to ensure, among other things, this constant updating, modification and revision. In fact, a 'monitor grammar' can presumably only be realised as an electronic grammar.

15 In this context, I regard it as a strength rather than a point of weakness of both CGEL and LGSWE not to refer explicitly to a particular grammatical framework in using the notions of 'canonical', 'usual', 'marked' and

‘unmarked’, because their analyses thus remain open to various models of grammar. This theoretical openness in itself is a major requirement that a reference grammar should meet (rather than try to force a specific theory of grammar on the user of a reference grammar). The ‘theoretical neutrality’ of CGEL does not mean, of course, that it is not based on specific descriptive and methodological principles as well as well-defined grammatical categories (see section 1), which are also adopted by LGSWE.

References

Aarts, F. (1975), ‘The Great Tradition or grammars and Quirk’s grammar’, in: *Dutch quarterly review of Anglo-American letters*, 5: 98-126.

Aarts, J. (2000), ‘Towards a new generation of corpus-based English grammars’, in: B. Lewandowska-Tomaszczyk and P.J. Melia (eds.) *PALC'99: practical applications in language corpora*. Frankfurt am Main: Peter Lang. 17-36.

Andersen, G. (2002), ‘Corpora and the double copula’, in: L.E. Breivik and A. Hasselgren (eds.) *From the COLT’s mouth …and others’: language corpora studies in honour of Anna-Brita Stenström*. Amsterdam: Rodopi. 43-58.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999), *Longman grammar of spoken and written English*. Harlow: Pearson Education. [LGSWE]

Brdar, M. (2002), ‘Yet another English reference grammar for the 21st century’, in: *The European English messenger*, 11(2): 79-82.

Carkin, S. (2000), ‘Review of D. Biber, S. Johansson, G. Leech, S. Conrad and E. Finegan, Longman grammar of spoken and written English (Harlow: Pearson Education, 1999)’, in: *Applied linguistics*, 21: 410-415.

Conrad, S. (2000), ‘Will corpus linguistics revolutionize grammar teaching in the 21st century?’, Paper read at the second North American symposium on corpora and language teaching, Flagstaff/AZ, 31.3.-2.4.2000.

Esser, J. (1992), ‘Neuere Tendenzen in der Grammatikschreibung des Englischen’, in: *Zeitschrift für Anglistik und Amerikanistik*, 40: 112-123.

Esser, J. (2002), ‘Sampling and categorizing fronted constructions in the BNC’, in: A. Fischer, G. Tottie and H.M. Lehmann (eds.) *Text types and corpora: studies in honour of Udo Fries*. Tübingen: Gunter Narr. 131-138.

Görlach, M. (2000), ‘Review of D. Biber, S. Johansson, G. Leech, S. Conrad and E. Finegan, Longman grammar of spoken and written English (Harlow: Pearson Education, 1999)’, in: *Arbeiten aus Anglistik und Amerikanistik*, 25: 257-260.

Huddleston, R. (1988), ‘Review of R. Quirk, S. Greenbaum, G. Leech and J. Svartvik, A comprehensive grammar of the English language (London: Longman, 1985)’, in: *Language*, 64: 345-354.

Huddleston, R. and G. K. Pullum (2002a), *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press. [CamGr]

Huddleston, R. and G. K. Pullum (2002b), 'Response to Joybrato Mukherjee regarding the Cambridge grammar of the English language', in: *Linguist list*, 13.1932, available at <http://www.linguistlist.org/issues/13/13-1932.html>.

Huddleston, R. and G. K. Pullum (2002c), 'Some points of agreement about the Cambridge grammar', in: *Linguist list*, 13.2005, available at http://www.linguistlist.org/issues/13/13-2005.html.

Huddleston, R. and G. K. Pullum (2003), 'English grammar', in: *The European English messenger*, 12(1): 65-67.

Hunston, S. (2002), *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Krug, M. (2002), 'Review of D. Biber, S. Johansson, G. Leech, S. Conrad and E. Finegan, Longman grammar of spoken and written English (Harlow: Pearson Education, 1999)', in: *English language and linguistics*, 6: 379-384.

Leech, G. (2004), 'A new Gray's anatomy of English grammar', in: *English language and linguistics,* 8: 121-147.

Mair, C. (1990), *Infinitival complement clauses in English: a study of syntax in discourse*. Cambridge: Cambridge University Press.

Mair, C. (2002), 'Three changing patterns of verb complementation in late modern English: a real-time study based on matching text corpora', *English language and linguistics*, 6: 105-131.

Mukherjee, J. (2002a), 'Review of R. Huddleston and G.K. Pullum, The Cambridge grammar of the English language (Cambridge: Cambridge University Press, 2002)', in: *Linguist list*, 13.1853, available at <http://www.linguistlist.org/issues/13/13-1853.html>.

Mukherjee, J. (2002b), 'A reply to Rodney Huddleston and Geoffrey K. Pullum concerning the Cambridge grammar of the English language', in: *Linguist list*, 13.1952, available at http://www.linguistlist.org/issues/13/13-1952.html.

Mukherjee, J. (2005), *English ditransitive verbs: aspects of theory, description and a usage-based model*. Amsterdam: Rodopi.

Parker, F. (2003), 'Review of D. Biber, S. Johansson, G. Leech, S. Conrad and E. Finegan, Longman grammar of spoken and written English (Harlow: Pearson Education, 1999)', in: *Journal of English linguistics*, 31: 90-97.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985), *A comprehensive grammar of the English language*. London: Longman. [CGEL]

Schmid, H.-J. (2003), 'Review of D. Biber, S. Johansson, G. Leech, S. Conrad and E. Finegan, Longman grammar of spoken and written English (Harlow: Pearson Education, 1999)', in: *Journal of pragmatics*, 35: 1265-1269.

Schneider, E.W. (2001): ‘Review of D. Biber, S. Johansson, G. Leech, S. Conrad and E. Finegan, Longman grammar of spoken and written English (Harlow: Pearson Education, 1999)’, in: *English world-wide*, 22: 137-143.

Sinclair, J. (ed.) (1987), *Collins COBUILD English language dictionary*. London: Collins.

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

Sinclair, J. (2002), ‘Review of D. Biber, S. Johansson, G. Leech, S. Conrad and E. Finegan, Longman grammar of spoken and written English (Harlow: Pearson Education, 1999)’, in: *International journal of corpus linguistics*, 6(2): 339-359.

Standop, E. (2000), ‘Englische Verbkomplementation’, in: *Anglia*, 118: 217-257.

Stubbs, M. (1993), ‘British traditions in text analysis: from Firth to Sinclair’, in: M. Baker, G. Francis and E. Tognini Bonelli (eds.) *Text and technology: in honour of John Sinclair*. Amsterdam: John Benjamins. 1-33.