16

What can a corpus tell us about grammar?

Susan Conrad

1 Understanding grammar through patterns and contexts: moving from correct/incorrect to likely/unlikely

In recent years, corpus-based studies of grammar have expanded greatly; new corpora, additional tools and more sophisticated analyses have all increased our understanding of grammars in different language varieties. Nonetheless, the essential, groundbreaking contribution of corpus linguistics to the study of grammar remains the same: Corpus linguistics changes our conceptualisation of grammar from a simple correct/incorrect dichotomy to an understanding of patterns and choices, an understanding of what is likely or unlikely in particular circumstances.

In traditional descriptions of grammar and in many syntactic theories, grammar is addressed through sample sentences that are either grammatical or ungrammatical, acceptable or unacceptable, accurate or inaccurate (e.g. see discussions in Cook 1994). From this perspective, to describe the grammar of a language, researchers need only to judge grammaticality, and to teach a language, instructors need only to focus on the rules for making grammatical sentences. Proficiency is equated with structural accuracy.

This dichotomous view works well for certain grammatical features. For example, it is grammatically incorrect to have zero article before a singular count noun in English: **I saw Ø cow*. Aside from a few exceptions such as in the locative prepositional phrases *at home* or *in hospital*, this rule is absolute for most varieties of English. However, any reflective language user will realize that many other grammatical choices cannot be made on the basis of correct/incorrect. For example, in the previous sentence the *that* could have been omitted: *…will realize many other grammatical choices….* Both versions are equally grammatical.

Of course, for decades, work in sociolinguistics and from a functional perspective has emphasised language choices for different contexts. In language classes, students may have been taught a few variants for politeness (e.g. in English using *could you* for requests instead of *can you*), but descriptions of grammar remained focused on accuracy. In a 1998 address to the international TESOL convention, Larsen-Freeman sought to 'challenge the common misperception that grammar has to do solely with

DOI: 10.4324/9780367076399-16

formal accuracy', arguing instead for a 'grammar of choice' (Larsen-Freeman 2002: 104). Being able to describe the typical choices that language users make, however, requires doing large-scale empirical analyses. The analyses must be empirical – rather than introspective – since language users may not be consciously aware of their choices. The analyses must cover numerous data in order to tell which language choices are widespread, which occur predictably although under rare circumstances and which are more idiosyncratic.

The great contribution of corpus linguistics to grammar is that it increases researchers' ability to systematically study the variation in a large collection of texts – produced by far more speakers and writers in more contexts than could be analyzed by hand. Corpus linguistic techniques allow us to determine common and uncommon choices and to see the patterns that reveal what is typical or unusual in particular contexts. These patterns show the correspondence between the use of a grammatical feature and some other factor in the discourse or situational context (e.g. another grammatical feature, a social relationship, the mode of communication, etc.). Corpus linguistics therefore allows us to focus on the patterns that characterise how a large number of people use the language, rather than basing generalisations on a small set of data or anecdotal evidence, or focusing on the accurate/inaccurate dichotomy. As O'Keeffe *et al.* (2007) explain, corpus analyses lead us to describing grammar not just in structural terms but in probabilistic terms.

This chapter reviews some major aspects of this corpus-based perspective for describing grammar. Section 2 reviews the types of grammatical patterns most typically covered in corpus studies, and Section 3 then discusses the investigation of numerous contextual factors simultaneously. Descriptions of grammar in spoken discourse are covered in Section 4, including some grammatical patterns specific to speech rather than writing. Section 5 concludes the chapter by reviewing some new challenges for the coming years. Throughout, the majority of references and examples refer to English. The contributions of corpus linguistics are equally applicable to the grammar of any language, but English continues to be the most-studied language. Furthermore, although numerous studies are mentioned, it is no coincidence that the chapter repeatedly cites two reference grammars of English: the *Grammar of Spoken and Written English* (Biber *et al.* 1999, 2021) and the *Cambridge Grammar of English* (Carter and McCarthy 2006). These comprehensive grammars make extensive use of corpus analyses to describe grammar structure and use, and they are currently the single clearest manifestations of corpus linguistics' impact on the study of grammar.

As a first step, before further discussing the contribution of corpus linguistics to grammar, a brief review of some methodological principles for corpus linguistic investigations of grammar is in order.

Methodological principles in corpus-based grammar analysis

Any analysis of "typical" or "probable" choices depends on frequency analysis. The very mention of a choice being typical or unusual implies that, under given circumstances, it happens more or less often than other choices. For reliable frequency analysis, a corpus does not always have to be immense, but it must be designed to represent a variety of language (see chapters in Part 1, this volume) and as fine-grained as needed to describe the circumstances associated with grammar choices. For example, Carter and McCarthy (2006: 11) find ellipses to be rare in narratives, while they are common in many other

parts of conversation. Any corpus that did not include numerous conversational genres or any analysis which neglected to differentiate among them would fail to discover this pattern.

Frequency counts are not sufficient for describing grammar, however. Instead, they point to interesting phenomena that deserve further investigation and interpretation. As Biber, Conrad and Cortes explain,

> …we do not regard frequency data as explanatory. In fact we would argue for the opposite: frequency data identifies patterns that must be explained. The usefulness of frequency data (and corpus analysis generally) is that it identifies patterns of use that otherwise often go unnoticed by researchers.
>
> *(Biber et al. 2004: 176)*

In corpus-based grammar studies, interpretations of frequency analyses come from a variety of sources. They can be based on cognitive principles such as the principle of end weight (heavy, long constituents are harder to process than short constituents and so are placed at the ends of clauses); on aspects of linguistic theory, such as principles defined in systemic functional linguistics; on the historical development of the language; or on reasonable explanations of the functions or discourse effect of a particular linguistic choice. Interpretation always includes human judgments of the impact of the language choices and speakers/writers' (usually subconscious) motivations in making these choices. Thus, a corpus linguistic perspective on grammar has not made human judgments superfluous; it has actually expanded the judgments and interpretations that are made.

2 Types of grammatical patterns

This section describes and exemplifies four types of patterns that are most common in corpus-based grammar analyses. Grammatical choices are associated with vocabulary, grammatical co-text, discourse-level factors and the context of the situation. Some other patterns specific to spoken discourse, such as associations with intonation, are covered in Section 4.

Grammar–vocabulary associations (lexico-grammar)

Associations between grammar and vocabulary are often called "lexico-grammar" or sometimes "colligation". The connection between words and grammar was extensively studied in the Collins COBUILD project (Sinclair 1991). Although designed initially as a lexicography project, it became clear that grammar and lexis were not as distinct as traditionally presented, and the project also resulted in a number of books presenting "pattern grammar" – explanations of grammatical structures integrated with the specific lexical items most commonly used in them (see Hunston and Francis 1999 and Chapter 11, this volume). Since that time, lexico-grammatical relationships of various sorts have been a common contribution of corpus studies.

One type of lexico-grammatical relationship concerns the lexical items that tend to occur with a particular grammatical structure. This type of pattern can be illustrated with verbs that are most common with *that*-clause objects, e.g. *I guess I should go* or *The results suggest that there is no effect.* A large number of verbs are possible with this

structure. However, beyond looking at what is possible, corpus-based grammar references present findings for the verbs that are actually most commonly used (Biber *et al.* 1999: 668–70; Carter and McCarthy 2006: 511). The reference grammars explain that the common verbs are related to expressing speech and thought. For example, Biber *et al.* (1999) find that *think, say* and *know* are by far the most common verbs with *that*-clauses in both British and American conversation (with the addition of *guess* in American English conversation). They also find that the structure is less common overall with any verb in academic prose, but *suggest* and *show* are most common. Rather than reporting thoughts and feelings, the verb + *that*-clause structures in academic prose are used to report previous research, often with non-human entities acting as the subject, as in *The results suggest that*….

This kind of lexico-grammatical pattern can also be approached from the perspective of the words themselves, usually for vocabulary teaching. The increasing use of computational analyses has expanded this area of work. For example, Ma and Qian (2020) use rule-based programming to find the most common grammar patterns for the most frequent verbs in academic English writing. They find, for instance, that the verb *require* is most common in a simple verb + noun pattern and second in a verb + noun + *to*-infinitive pattern.

Another type of lexico-grammatical relationship concerns the specific words that occur as a realisation of a grammatical function. A simple illustration is verb tense. Traditionally, a grammatical description would simply explain the form of tenses – for example, that simple present tense in English is uninflected except in third person singular when *-s* is added, that past tense is formed with *-ed* for regular verbs, etc. A corpus-based grammar can add information about the verbs used most commonly in the tenses. For example, in the *Longman Spoken and Written English Corpus*, the set of verbs occurring over 80 per cent of the time in present tense differs greatly from the verbs occurring over 80 per cent of the time in past tense (Biber *et al.* 1999: 25). The verbs common in present tense convey mental, emotional and logical states. Many are used in short, common expressions in conversations expressing the speaker's thoughts or feelings, such as *It doesn't matter* or *Never mind,* while others are used to describe the states of others or to make logical interpretations about what something *means* or what someone *doubts*. The verbs more strongly associated with past tense, on the other hand, convey events or activities, especially body movements and speech (e.g. *exclaimed, glanced, grinned, nodded, whispered*). Not surprisingly, those past tense verbs are especially common for describing characters and actions in fiction writing.

The associations between a grammatical structure and lexical items can also be analysed in terms of semantic characteristics, leading to an analysis of "semantic prosody" – the fact that certain structures tend to be associated with certain types of meaning, such as positive or negative circumstances (Sinclair 1991; Louw 1993). O'Keeffe *et al.* (2007: 106–14) provide an analysis of *get*-passives (e.g. *he got arrested*). They show that the *get*-passive is usually used to express unfortunate incidences, manifest in the lexico-grammatical association of verbs such as *killed, sued, beaten, arrested, burgled, intimidated, criticized* and numerous others. Stempel (2019) shows that the "into Ving causative" structure – e.g. *talked into buying, fooled into thinking* – is used to express negative attitudes towards the proposition, often with coercion or deception involved. These authors and others investigating semantic prosody note also that individual lexical items may not be negative – there is nothing inherently negative about *talked*, for example – but the discourse context makes the adverse connections clear.

O'Keeffe *et al.* (2007) also discuss the type of subjects usually found with *get*-passives (often human subjects – the people to whom the unfortunate incident happened) and the lack of adverbials in these clauses. The authors thus move into discussion of another type of pattern, the grammatical co-text.

Grammatical co-text

Corpus studies often investigate the extent to which a particular grammatical feature tends to occur with specific other grammatical features. Grammatical descriptions in traditional textbooks sometimes make claims about the grammatical co-text of features, and corpus studies can provide empirical testing of these claims. An interesting example is provided by Frazier (2003), who investigated *would*-clauses of hypothetical or counterfactual conditionals. Concerned about the way that English as second language (ESL) grammars virtually always present the *would*-clause as adjacent to an *if*-clause, he examined the extent to which this was true in a combination of spoken and written corpora.

Surprisingly, Frazier (2003) found that almost 80 per cent of the hypothetical/counterfactual *would*-clauses were not adjacent to an *if*-clause. Although there are several categories of *would*-clauses without *if*-clauses, the largest percentage were those that had implied, covert conditionals. It further turned out that these clauses tended to occur with certain other grammatical features, including infinitives and gerunds, as in these examples:

> If there is nothing evil in these things, if they get their moral complexion only from our feeling about them, why shouldn't they be greeted with a cheer? *To greet* them with repulsion *would* turn what before was neutral into something bad…
>
> *Letting* the administration take details off their hands *would* give them more time to inform themselves about education as a whole…
>
> *(Frazier 2003: 456–7)*

More recently, studies of the use of *if*-conditionals have expanded to include modality marking more generally and provide more theoretical explanations for patterns (Gabrielatos 2019).

Looking at grammatical co-occurrence patterns can also help to explain when rare constructions occur. For example, subject position *that*-clauses, as illustrated here, are very rare:

> (1)
> That there are no meteorites of any other age, regardless of when they fell to Earth suggests strongly that all meteorites originated in other bodies of the solar system that formed at the same time that the Earth did.
>
> *[Longman Spoken and Written English Corpus]*

Considering constructions with *that* and *the fact that*, subject position clauses occur about 20 to 40 times per million words in academic prose and newspapers and almost never occur in conversation, while *that*-clauses in other positions occur over 2,000 to 7,000 times per million words in the different registers (Biber *et al.* 1999: 674–6). These

subject position clauses are obviously harder for listeners or readers to process, since they have a long constituent before the main verb. It is perhaps not surprising, then, that the subject position clauses tend to occur when the predicate of the sentence has another heavy, complex structure – a complicated noun phrase or prepositional phrase, or a complement clause, as in (1). In addition, these clauses tend to be used in particular discourse contexts, a topic further discussed in the next section.

Discourse-level factors

Many people's introduction to corpus linguistics is with simple concordance searches or collocate lists, and they sometimes believe that corpus linguistics has little to offer discourse-level study. However, this clearly is not the case (see Chapter 18, this volume).

Analysis of discourse-level factors affecting grammar often requires interpreting meaning, organisation and information structure in texts. Such analysis is part of the more qualitative, interpretive side of a corpus study, focusing on how a grammatical structure is used in context. In fact, several examples in the previous sections have noted associations that were found by analysing text at the discourse level rather than considering only discrete lexical or grammatical features. Determining semantic prosody, for instance, requires considering discourse context and assessing if it is negative.

The status of information is often important in grammar-discourse association patterns. Most commonly, studies find patterns related to whether information has already been mentioned in the discourse or is new. For example, the rare subject position *that*-clauses described in the previous section tend to restate information that has already been mentioned or implied in the previous discourse, so the subject clauses provide an anaphoric link. Similarly, Prado-Alonso (2019: 30) finds that subject-auxiliary inversion in declarative clauses (e.g. *Never have I*) serves a "focus management" function, working new information into the discourse.

Another perspective on grammar and discourse concerns matching grammar choices to their rhetorical functions. Williams (2010), for example, finds that first person pronouns tend to occur in English medical research articles in particular rhetorical moves, where researchers explain their choice of non-standard procedures, but in Spanish articles, first person is used in different rhetorical moves and with both non-standard and standard procedures, thereby creating empathy for the doctor–patient context. Studying social science research article introductions, Lu *et al.* (2020) analyse several measures of syntactic complexity and their associations with rhetorical moves. They find that longer sentences with higher density nominalisations and more frequent nonfinite dependent clauses are associated with the rhetorical move of announcing the present research.

Grammatical complexity in different types of discourse has also received a great deal of attention. Much of this work can be found in analyses of register (see Chapter 17, this volume). A major distinction found by studies has been the tendency for academic prose to have phrasal complexity, while conversation tends to have clausal complexity (Biber *et al.* 2011; Biber and Gray 2016). Research articles written in an English as a lingua franca context have been found to use complex nominal structures, too, though with some other syntactic complexity differences from American English research article writing (Wu *et al.* 2020).

Context of the situation

A number of factors in the context of the situation may be associated with the choice of a particular grammatical feature. Sociolinguistic studies have long considered how language use is affected by audience, purpose, participant roles, formality of the situation and numerous other social and regional characteristics, and corpus-based techniques can be applied in these areas. Thus far in studies of grammar, the most common perspective on variation has concerned registers (also called genres) – varieties associated with a particular situation of use and communicative purpose, and often identified within a culture by a specific name, such as academic prose, text messaging, conversation or newspaper writing. Comparisons of grammar in major registers such as academic writing and conversation have been included in descriptions of grammar studies in the previous sections. Other studies compare registers in more restricted domains. Grammatical features used in academic settings have received considerable attention. Numerous studies have analysed grammar features in universities – e.g. Fortanet (2004) describes details of the pronoun *we* in lectures, Louwerse *et al.* (2008) discuss the use of conditionals and Biber (2006, chapter 4) compares numerous grammatical features across ten spoken and written registers. A few studies compare academic settings to other professional settings. For example, in a study of engineering, Conrad (2018) finds a lower frequency of passives and other impersonal style features in writing by industry professionals compared to research journal articles, tying the lower frequencies to concerns with agency, conciseness and ease of reading for clients.

Whether focused on general or more specific settings, studies that make comparisons across registers all demonstrate that it is usually misleading to characterise the frequency and use of a grammatical feature in only one way. Instead, accurate grammatical descriptions require describing differences across registers (see Conrad 2000 and Chapter 17, this volume).

Traditional sociolinguistic variables having to do with social groups and regions were usually addressed in general categories in early corpus-based grammar research, with categories such as British and American English (e.g. in various comparisons throughout Biber *et al.* 1999; Carter and McCarthy 2006 appendix), varieties of world Englishes (Kachru 2008) or a general group such as London teenagers (Stenström *et al.* 2002). While these general classifications are still a popular topic (e.g. Baker 2017), more refined variables are becoming increasingly common in corpus studies. Calude (2017) includes gender, age, education and occupation as variables in investigating the use of demonstrative cleft constructions (such as *This is where I saw him*), finding the clefts are more typical of male rather than female speakers, adults in higher skilled jobs as compared to adults in semi-skilled jobs and middle-aged speakers over younger speakers. There are also increasingly specific regional dialect categories, as with Fernandez-Ordonez's (2010) study of grammar in rural Spanish dialects.

3 Investigating multiple influences on grammatical patterns

From the previous sections, it is probably already apparent that it is often difficult to focus on only one type of pattern when explaining grammatical choices. However, before computer-assisted analyses, it was unfeasible to consider multiple factors in a large number of texts simultaneously. Another contribution of corpus linguistics, then, has been to describe more about the multiple factors that simultaneously have an impact on

grammatical choices. Corpus-based studies have used several approaches for examining multiple factors, depending on the goals and audience for the work.

Some studies provide a description of multiple factors without statistical analysis. For example, consider the case of omitting the optional *that* in a *that*-complement clause – e.g. *I think Ø I'll go*. Virtually any grammatical description includes the fact that the *that* is optional. Textbooks for ESL students often explain that it is especially common to delete it in speech (e.g. Azar 2002: 248), but a corpus analysis reveals that omission of *that* is actually associated with a number of factors (Biber *et al.* 1999: 681), as shown in Figure 16.1. One factor is a lexico-grammatical association: *That* is omitted more often when the verb in the main clause is *say* or *think* rather than any other verb. Two factors concern the grammatical co-text: *That* is omitted more often when (1) the main clause and complement clause have co-referential subjects rather than subjects that refer to different entities or (2) the *that*-clause has a personal pronoun subject rather than a full noun phrase. Another factor concerns the situational context: *That* is omitted more often in conversation than in newspaper writing generally, but the lexico-grammatical and grammatical co-text factors have a stronger effect in newspapers. That is, the choice of verb and subject types corresponds to a greater difference in percentage of *that* omission in newspapers than in conversation.

Other studies rely more on statistical analysis, which allows researchers to understand the relationship between factors in more detail. For example, Schilk *et al.* (2013) use multinomial logistic regression in studying the complementation patterns of the verb *give* in web-derived corpora of Indian, Pakistani and British English (e.g. *She gave the class a presentation* vs. *She gave a presentation to the class* vs. *She gave a presentation)*. They find three variables that affect the choice to differing extents: pronoun vs. noun use, distance since participants were mentioned in the discourse and the language variety. A statistical approach is becoming increasingly common in studies of grammar choices, ranging, for example, from the choice of indicative and subjunctive alternation in subordinate clauses in Spanish (Deshors and Waltermire 2019) to the choice of topic marking in spoken Shanghainese (Han *et al.* 2017).

Two other approaches are also often used for analysing multiple influences on grammatical choices. One is to consider a functional system within a language variety

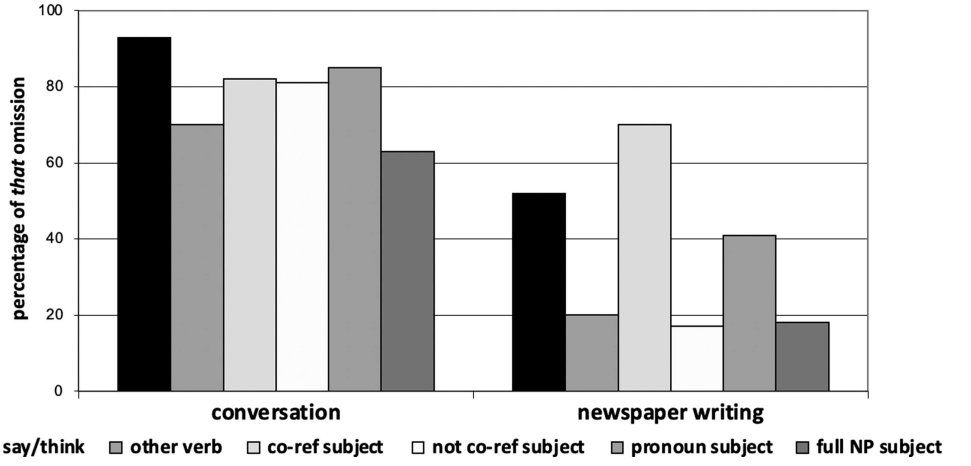

Figure 16.1 Conditions associated with omission of *that* in *that*-clauses

and describe factors that influence the grammatical features that are used to realise the system. For example, the system of metadiscourse or stance is often studied; Hyland (2017) provides a review of studies of the realisation of metadiscourse in many registers, from Twitter to research articles, most of them using a corpus-based lexico-grammatical perspective, and Gales (2015) examines the grammatical marking of stance in written threats. The other approach is to study the grammar of a variety. In this approach the focus shifts from describing grammar to describing the variety, as covered by Gray (Chapter 17, this volume).

4 Grammar and speech

The previous sections have all made mention of grammatical features in spoken discourse. Before corpus studies became popular, grammatical descriptions were based on written language (McCarthy and Carter 1995; Carter and McCarthy 2017). Unplanned spoken language was neglected or, at best, considered aberrant, with incomplete clauses, messy repairs and non-standard forms. In contrast, corpus analyses have emphasised the fact that many features of speech directly reflect the demands of social interactions. Grammar in speech, especially in conversation, has become studied as a legitimate grammar, not a lacking form of written grammar.

One factor often noted for grammatical choices in conversation concerns the need to minimise imposition and be indirect. For example, Conrad (1999) discusses the common choice of *though* rather than *however* as a contrastive connector in conversation. Placed at the end of the clause and conveying a sense of concession more than contrast, the use of *though* is a less direct way to disagree than *however*. A typical example in a conversation is as follows:

(2)
[Watching a football game, discussing a penalty call]
A:
Oh, that's outrageous.
B:
Well, he did put his foot out **though**.
[Longman Spoken and Written English Corpus]

Speaker B clearly disagrees with A's contention that the call is "outrageous", but the use of *though* (along with the discourse marker *well*) downplays the disagreement. The desire to convey indirectness can also result in the use of verb tenses or aspects not typically found in writing. For example, McCarthy and Carter (2002: 58) describe the use of present progressive with verbs of desire, as when a customer tells a travel agent that she and her husband *are wanting* to take a trip. The use of present progressive makes the desire sound more tentative and the request for help less imposing.

Other features affecting the grammatical forms that are typical of face-to-face interactions include the shared context (and in many cases shared background knowledge), the expression of emotions and evaluations and the constraints of real-time production and processing (see a further summary of factors and features in Biber *et al.* 1999 chapter 14, and Carter and McCarthy 2006: 163–75). Some of these conditions lead to common use of grammar features that are typically not even mentioned

in grammars based on writing. For instance, work on discourse markers (e.g. *well, you know, I mean*) has flourished in corpus-based studies, not just of English but numerous languages, including multi-word discourse markers in Slovene (Dobrovoljc 2017) and the co-occurrence of discourse markers and disfluencies in French and English (Crible 2017).

Corpus-based studies of grammar in spoken discourse have also found associations between grammatical structures and intonation. While corpus-based work with intonation has been going on since the 1980s when the London-Lund Corpus was released with basic intonation in the transcription, the role of intonation in disambiguating the functions of features has increased in recent years. Work has included clausal structures; for example, Adolphs and Carter (2013) note that it is possible to disambiguate use of *I don't know why* as a full clause or sentence stem by considering intonation unit boundaries. Studies have also covered specific words that can realise different grammatical classes or classes that have multiple meanings. For example, Wichmann *et al.* (2010) find that *of course* is marked with prosodic prominence when it functions as an epistemic stance adverbial but not when it is a more general discourse marker. Biber and Staples (2014) find that different prosodic patterns are associated with different functions for high-frequencystance adverbials such as *actually*; when a speaker emphasises the semantic content, such as *actually* meaning "in actual fact", the adverbials are more likely to have prosodic prominence, but when the adverbials express general certainty or doubt, prosodic prominence is less likely.

In addition to intonation, many spoken interactions are distinct from writing by having a shared visual component. Exploring the patterns between grammar and non-verbals is one of the new challenges discussed in the next section.

5 New challenges with grammar and corpora

The first edition of this chapter (Conrad 2010) concluded by discussing two controversies over the value of corpus-based grammar study. The first controversy focused on teaching and learning, especially the argument that corpus investigations could be harmful because corpora would show learners unusual constructions in addition to common ones (see Owen 1996 and the response in Hunston 2002). Ten years later, this controversy has faded. Although there are many considerations for applications to language teaching, corpus linguistics has become a more firmly established way to explain grammatical choices (e.g. see Chapters 25 and 30, this volume).

A second controversy ten years ago was whether evidence from corpus studies of grammar could contribute to linguistic theory (e.g. Newmeyer 2003; Meyer and Tao 2005). Today it is clear that such evidence has a role to play. Not surprisingly, many corpus-based studies are associated with construction grammar (e.g. Gabrielatos 2019; Wible and Tsao 2020), which views multi-word patterns as building blocks of syntax, not unlike pattern grammar (Hunston 2019). Other studies contribute to the development of theories that, at first, might appear at odds with corpus techniques. Studies applying systemic functional linguistics (SFL), for instance, traditionally tended to present short text examples, but integration of SFL theory with corpus analysis is now common. Some studies focus on specific grammatical features, such as Xiang and Liu (2018)'s study of *let's* constructions, clarifying the system of MOOD in SFL. Others focus on more general constructions, such as Lee's (2016) study of South Korean newspaper reports, which expands on SFL transitivity analysis.

Today the challenges for corpus-based grammar studies hinge not so much on basic questions of whether corpus analysis is valuable, but rather on increasing the depth of analysis. As noted in the previous section, one new challenge is multi-modal analysis, investigating how patterns of grammar correspond with gestures, facial expressions and body movements (see Chapter 7, this volume). Though more work thus far has focused on words and phraseology, some includes grammar, such as identifying the grammatical categories that tend to be associated with gestures and those that do not (Kok 2017). Multi-modal analysis, however, presents several conundrums for an approach whose strengths include naturally occurring contexts and large databases of text. In addition to most people finding video recording more intrusive than audio recording, the analysis of eye gaze, facial expressions and small movements requires high-quality video. It is thus currently unfeasible to create a multi-modal corpus that includes all kinds of naturally occurring spoken interactions. In addition, analysis of non-verbals is time-consuming. Software for the coding of non-verbals is helpful (see Adolphs and Carter 2013 and Chapter 7, this volume), but analysis simply cannot cover as many speakers or as much language as in text-only analyses. The depth of analysis requires a trade-off in the breadth of texts.

The desire for more interpretive depth in corpus studies is apparent in some other developments as well. Numerous studies now investigate lexico-grammatical associations as part of critical discourse analysis (CDA), with the corpus analysis showing the systematic patterns of language use and the CDA providing more about the features' connections to groups' representations, identities and power dynamics (e.g. Wilkinson 2019 on the representation of bisexuals; Alcantud-Díaz, 2012 on the identity, social power and violence in Grimm's fairy tales; Potts *et al.* 2015 on the linguistic construction of newsworthiness for a hurricane). Nartley and Mwinlaaru (2019) provide a meta-analysis of 121 studies combining corpus linguistics and CDA. Corpus studies that cover some grammar features and lexico-grammar have also been combined with conversation analysis (Walsh 2013), so the context of the features and their use in interaction can be analysed more closely than through corpus techniques alone. Other studies advocate incorporating more input from speakers or writers represented in a corpus in order to understand the grammar patterns more fully within their contexts of use; Conrad (2021), for example, argues for the need for writing studies to combine corpus analysis and interviews of writers to understand the intentions behind their grammar choices and create targeted instructional materials. For all these approaches, balancing the strength of large-scale corpus analysis while adding more intensive analytical techniques presents a challenge that continues to be refined.

Unfortunately, a chapter of this size cannot do justice to many aspects of corpus-based grammar analysis. In fact, even traditional grammar terminology used throughout this chapter deserves interrogation, as corpus researchers highlight the gaps between traditional meta-language and features of speech and electronic communication (Carter and McCarthy 2017). Further attention is due also to diachronic studies that explain changes in grammar use over time (e.g. Biber and Gray 2016; Jenset and McGillivray 2017) and to many specific grammar features of language varieties in the world (e.g. Esimaje *et al.* 2019). Other chapters in this volume add important perspectives on methodological issues (Part I, this volume) and teaching (Part III, this volume). Nonetheless, despite its inability to do justice to all related topics, the chapter has shown that corpus linguistics has already had a profound effect on our understanding of grammar and is likely to continue to do so in the future.

Further reading

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (2021) *Grammar of Spoken and Written English*, Amsterdam: John Benjamins. (This reference grammar covers frequency information, lexico-grammar patterns and comparisons of use in conversation, fiction writing, newspaper writing and academic prose for the major structures in English, in addition to chapters on fixed phrases, stance and conversation. It includes everything from the earlier *Longman Grammar of Spoken and Written English.*)

Carter, R. A. and McCarthy, M. J. (2006) *Cambridge Grammar of English*, Cambridge: Cambridge University Press. (This reference grammar emphasises spoken vs. written language for major grammar features of English and includes many lexico-grammatical analyses. It also covers some functional categories, such as typical grammatical realisations of speech acts and many typical ESL difficulties.)

References

Adolphs, S. and Carter, R. A. (2013) *Spoken Corpus Linguistics: From Monomodal to Multimodal*, New York: Routledge.

Alcantud-Díaz, M. (2012) 'The Sister Did Her Every Imaginable Injury: Power and Violence in Cinderella', *International Journal of English Studies* 12(2): 59–71.

Azar, B. (2002) *Understanding and Using English Grammar*, 3rd edn, White Plains, NY: Longman.

Baker, P. (2017) *American and British English: Divided by a Common Language?* Cambridge: Cambridge University Press.

Biber, D. (2006) *University Language: A Corpus-Based Study of Spoken and Written Registers*, Amsterdam: John Benjamins.

Biber, D. and Gray, B. (2016) *Grammatical Complexity in Academic English: Linguistic Change in Writing*, Cambridge: Cambridge University Press.

Biber, D. and Staples, S. (2014) 'Exploring the Prosody of Stance: Variation in the Realization of Stance Adverbials', in T. Raso and H. Mello (eds) *Spoken Corpora and Linguistic Studies*, Amsterdam: John Benjamins, pp. 271–94.

Biber, D., Conrad, S. and Cortes, V. (2004) ' "Take a Look At…": Lexical bundles in University Teaching and Textbooks', *Applied Linguistics* 25(3): 401–35.

Biber, D., Gray, B. and Poonpon, K. (2011) 'Should We Use Characteristics of Conversation to Measure Grammatical Complexity in L2 Writing Development?', *TESOL Quarterly* 45(1): 5–35.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English*, Harlow, England: Pearson Education.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (2021) *Grammar of Spoken and Written English*, Amsterdam: John Benjamins.

Calude, A. (2017) 'Sociolinguistic Variation at the Grammatical/Discourse Level: Demonstrative Clefts in Spoken British English', *International Journal of Corpus Linguistics* 22(3): 429–55.

Carter, R. A. and McCarthy, M. J. (2006) *Cambridge Grammar of English*, Cambridge: Cambridge University Press.

Carter, R. A. and McCarthy, M. J. (2017) 'Spoken Grammar: Where Are We and Where Are We Going?' *Applied Linguistics* 38(1): 1–20.

Conrad, S. (1999) 'The Importance of Corpus-Based Research for Language Teachers', *System* 27(1): 1–18.

Conrad, S. (2000) 'Will Corpus Linguistics Revolutionize Grammar Teaching in the 21st Century?' *TESOL Quarterly* 34(3): 548–60.

Conrad, S. (2010) 'What Can a Corpus Tell Us about Grammar?', in A. O'Keeffe and M. J. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*, London: Routledge, pp. 227–240.

Conrad, S. (2018) 'The Use of Passives and Impersonal Style in Civil Engineering Writing', *Journal of Business and Technical Communication* 32(1): 38–76.

Conrad, S. (2021) 'Integrating Corpus Linguistics into Writing Studies: An Example from Engineering', in K. Blewett, T. Donahue and C. Monroe (eds) *The Expanding Universe of Writing Studies: Higher Education Writing Research*, New York: Peter Lang, pp. 43–56.
Cook, V. (1994) 'Universal Grammar and the Learning and Teaching of Second Languages', in T. Odlin (ed.) *Perspectives on Pedagogical Grammar*, Cambridge: Cambridge University Press, pp. 25–48.
Crible, L. (2017) 'Discourse Markers and (Dis)fluency in English and French', *International Journal of Corpus Linguistics* 22(2): 242–69.
Deshors, S. and Waltermire, M. (2019) 'The Indicative vs. Subjunctive Alternation with Expressions of Possibility in Spanish: A Multifactorial Analysis', *International Journal of Corpus Linguistics* 24(1): 67–97.
Dobrovoljc, K. (2017) 'Multi-Word Discourse Markers and Their Corpus-Driven Identification: The Case of MWDM Extraction from the Reference Corpus of Spoken Slovene', *International Journal of Corpus Linguistics* 22(4): 551–82.
Esimaje, A., Gut, U. and Antia, B. (eds) (2019) *Corpus Linguistics and African Englishes*, Amsterdam: John Benjamins.
Fernandez-Ordonez, I. (2010) 'Investigating Spanish Dialectal Grammar with the COSER (Audio Corpus of Spoken Rural Spanish)', *Corpus* 9: 81–114.
Fortanet, I. (2004) 'The Use of "We" in University Lectures: Reference and Function', *English for Specific Purposes* 23(1): 45–66.
Frazier, S. (2003) 'A Corpus Analysis of *Would*-Clauses Without Adjacent *If*-Clauses', *TESOL Quarterly* 37(3): 443–66.
Gabrielatos, C. (2019) 'If-Conditions and Modality: Frequency Patterns and Theoretical Explanations', *Journal of English Linguistics* 47(4): 301–34.
Gales, T. (2015) 'The Stance of Stalking: A Corpus-Based Analysis of Grammatical Markers of Stance in Threatening Communications', *Corpora* 10(2): 171–200.
Han, W., Arppe, A. and Newman, J. (2017) 'Topic Marking in a Shanghainese Corpus: From Observation to Prediction', *Corpus Linguistics and Linguistic Theory* 13(2): 291–19.
Hunston, S. (2002) *Corpora in Applied Linguistics*, Cambridge: Cambridge University Press.
Hunston, S. (2019) 'Patterns, Constructions and Applied Linguistics', *International Journal of Corpus Linguistics* 24(3): 324–53.
Hunston, S. and Francis, G. (1999) *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*, Amsterdam: Benjamins.
Hyland, K. (2017) 'Metadiscourse: What Is It and Where Is It going?' *Journal of Pragmatics* 113: 16–29.
Jenset, G. and McGillivray, B. (2017) *Quantitative Historical Linguistics: A Corpus Framework*, Oxford: Oxford University Press.
Kachru, Y. (2008) 'Language Variation and Corpus Linguistics', *World Englishes* 27(1): 1–8.
Kok, K. (2017) 'Functional and Temporal Relations between Spoken and Gestured Components of Language: A Corpus-Based Inquiry', *International Journal of Corpus Linguistics* 22(1): 1–26.
Larsen-Freeman, D. (2002) 'The Grammar of Choice', in E. Hinkel and S. Fotos (eds) *New Perspectives on Grammar Teaching in Second Language Classrooms*, Mahwah, NJ: Erlbaum, pp. 103–18.
Lee, C. (2016) 'A Corpus-Based Approach to Transitivity Analysis at Grammatical and Conceptual Levels', *International Journal of Corpus Linguistics* 21(4): 465–98.
Louw, B. (1993) 'Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies', in M. Baker, G. Francis, and E. Tognini-Bonelli (eds) *Text and Technology: In Honour of John Sinclair*, Philadelphia/Amsterdam: John Benjamins, pp. 157–76.
Louwerse, M., Crossley, S. and Jeuniaux, P. (2008) 'What If? Conditionals in Educational Registers', *Linguistics and Education* 19(1): 56–69.
Lu, X., Casal, J. and Liu, Y. (2020) 'The Rhetorical Functions of Syntactically Complex Sentences in Social Science Research Article Introductions', *Journal of English for Academic Purposes* 44: 1–16.
Ma, H. and Qian, M. (2020) 'The Creation and Evaluation of a Grammar Pattern List for the Most Frequent Academic Verbs', *English for Specific Purposes* 58: 155–69.
McCarthy, M. J. and Carter, R. A. (1995) 'Spoken Grammar: What Is It and How Do We Teach It?' *ELT Journal* 49(3): 207–18.

McCarthy, M. J. and Carter, R. A. (2002) 'Ten Criteria for a Spoken Grammar', in E. Hinkel and S. Fotos (eds) *New Perspectives on Grammar Teaching in Second Language Classrooms*, Mahwah, NJ: Lawrence Erlbaum, pp. 51–75.

Meyer, C. and Tao, H. (2005) 'Response to Newmeyer's "Grammar Is Grammar and Usage Is Usage"', *Language* 81(1): 226–8.

Nartley, M. and Mwinlaaru, I. (2019) 'Towards a Decade of Synergising Corpus Linguistics and Critical Discourse Analysis: A Meta-Analysis', *Corpora* 14(2): 203–35.

Newmeyer, F. (2003) 'Grammar is Grammar and Usage is Usage', *Language* 79(4): 682–707.

O'Keeffe, A., McCarthy, M. J. and Carter, R. A. (2007) *From Corpus to Classroom: Language Use and Language Teaching*, Cambridge: Cambridge University Press.

Owen, C. (1996) 'Does a Corpus Require to Be Consulted?' *ELT Journal* 50(3): 219–24.

Potts, A., Bednarek, M. and Caples, H. (2015) 'How Can Computer-Based Methods Help Researchers to Investigate News Values in Large Datasets? A Corpus Linguistic Study of the Construction of Newsworthiness in the Reporting on Hurricane Katrina', *Discourse & Communication* 9(2): 149–72.

Prado-Alonso, C. (2019) 'A Comprehensive Corpus-Based Analysis of "X Auxiliary Subject' Constructions in Written and Spoken English', *Topics in Linguistics* 20(2): 17–32.

Schilk, M., Mukherjee, J., Nam, C., and Mukherjee, S. (2013) 'Complementation of Ditransitive Verbs in South Asian Englishes: A Multifactorial Analysis', *Corpus Linguistics and Linguistic Theory* 9(2): 187–225.

Sinclair, J. (1991) *Corpus Concordance Collocation*, Oxford: Oxford University Press.

Stempel, P. (2019) 'A Constructional Reanalysis of Semantic Prosody', unpublished Ph.D. dissertation, Rice University.

Stenström, A., Andersen, G. and Hasund, I. (2002) *Trends in Teenage Talk: Corpus Compilation, Analysis and Findings*, Amsterdam: John Benjamins.

Walsh, S. (2013) 'Corpus Linguistics and Conversation Analysis at the Interface: Theoretical Perspectives, Practical Outcomes', in J. Romero-Trillo (ed.) *Yearbook of Corpus Linguistics and Pragmatics*, Dordrecht: Springer, pp. 35–51.

Wible, D. and Tsao, N. (2020) 'Constructions and the Problem of Discovery: A Case for the Paradigmatic', *Corpus Linguistics and Linguistic Theory* 16(1): 67–93.

Wichmann, A., Simon-Vandenbergen, A. and Aijmer, K. (2010) 'How Prosody Reflects Semantic Change: A Synchronic Case Study of *Of Course*', in K. Davidse, L. Vandelanotte and H. Cuyckens (eds) *Subjectification, Intersubjectification and Grammaticalization*, Berlin: Walter De Guyter, pp. 103–54.

Wilkinson, M. (2019) '"Bisexual Oysters": A Diachronic Corpus-Based Critical Discourse Analysis of Bisexual Representation in *The Times* between 1957 and 2017', *Discourse & Communication* 13(2): 249–67.

Williams, I. (2010) 'Cultural Differences in Academic Discourse: Evidence from First-Person Verb Use in the Methods Sections of Medical Research Articles', *International Journal of Corpus Linguistic* 15(2): 214–39.

Wu, X., Mauranen, A., and Lei, L. (2020) 'Syntactic Complexity in English as a Lingua Franca Academic Writing', *Journal of English for Academic Purposes* 43: 1–13.

Xiang, D. and Liu, C. (2018) 'The Semantics of mood and the Syntax of the Let's-Construction in English: A Corpus-Based Cardiff Grammar Approach', *Australian Journal of Linguistics* 38(4): 549–85.