9 Conclusion

9.1 Introduction

The preceding chapter concluded the survey of corpus methods in different fields of linguistics with which the latter part of this book has mainly been occupied. We have looked at the intersection of corpus methodologies with areas such as discourse analysis, sociolinguistics, language change, functionalist linguistics and psycholinguistics. In this final chapter, we will reflect on what we can conclude about the status of corpus linguistics within linguistics – looking at trends evident in the history of corpus linguistics up to the present time and considering how those trends are likely to continue, or, rather, how we think they should continue. In particular we will consider the future of corpus analysis within a framework of methodological pluralism, and the potential for corpus methods to extend beyond the field of linguistics into other areas of the humanities, sciences and social sciences. How, for example, can the methods and findings of corpus linguistics and computational linguistics continue to usefully interact? How can corpus methods be utilised in the analysis of the textually mediated world found in humanities subjects such as history, literary criticism and religious studies? And how can new methods in linguistics – for example, new approaches to neurolinguistics – inform the findings of corpus-based analyses through the process of methodological triangulation?

9.2 The story of corpus linguistics, from past to future

By surveying the variety of approaches to, and applications of, corpus linguistics over the past forty to fifty years, we have presented what may be called the ‘story’ of corpus linguistics. But what is the overarching theme of this narrative? In our view, essentially two broad phases in the history of corpus linguistics may be observed. The first stage, up to about the end of the 1980s, centres around the emergence of corpus linguistics primarily within two different schools of English language studies, its struggle to establish itself in the face of Chomskyan views inherently opposed to the use of corpora, and the formation of the basic set of methods and tools. The theme of the second phase, from that point
up until the present day, has been the shift in the nature of corpus linguistics as an enterprise that we have outlined in the latter part of this book. From being in practice a semi-independent subfield of linguistics – whether considered one in theory by its practitioners or not – corpus linguistics has become an indispensable component of the methodological toolbox throughout linguistics. The subfield labelled corpus linguistics that could have been coherently argued to exist in, say, 1990, is no longer so easily distinguishable from other forms of linguistics – and as we argued in the previous chapter, this rapprochement is even observable for the subset of neo-Firthian corpus linguists who do attempt to distinguish corpus linguistics as a separate field of linguistics.

Will there be a distinctive third phase in the ongoing development of corpus linguistics? Or, to put it more prosaically, what kind of future progressions can we predict for corpus linguistics? We anticipate that the trend of convergence between corpus linguistics and other types of linguistics will continue. The logical end-point of this development would be the extinction of corpus linguistics as a separate enterprise – that is, a situation where corpus methods are simply used (where appropriate) by all linguists rather than being the preserve of a marginalised subgroup, as was arguably the case up until the 1990s. We do not think that this end-point can (or should) ever be reached, because even when corpus methods are fully embedded in the day-to-day practice of functionalist linguistics, sociolinguistics, discourse analysis and so on, there will still be an important role for corpus specialists whose research is concerned with the methodology itself – the construction and annotation of corpora, the development of new tools and new procedures, the expansion of the conceptual bases of the methodology and other such issues. The highly technical nature of some aspects of corpus methods means that not all users can be thorough specialists in the methodology; as we have noted (in Chapter 2) we do not think it is realistic to expect every linguist who uses corpus data to become fully competent in computer programming, for instance, or in the more complex statistical analyses. So we may expect a shift in what it means to be a ‘corpus linguist’ – from meaning someone who uses corpus data in their research, to meaning a researcher into the methodology, especially one who develops new methods and enables other linguists to apply them. To an extent this shift has already taken place, in that the research with the greatest impact in corpus linguistics is very often valued not for what it discovers about language, but for the methods it introduces or develops. To take a recent example, the findings about particular English grammatical constructions made by Stefanowitsch and Gries (2003), which we discussed in Chapter 7, are not especially revolutionary in themselves. It is, rather, the method that these findings exemplify – and the associated theoretical and statistical apparatus linked to collostruction – that makes this paper a key contribution to recent research in corpus linguistics. Looking further back, Sinclair’s seminal (1991) text is cited more often for the illustration it gives of one prominent approach to corpus analysis than for the specific findings Sinclair presents in that work about particular linguistic items.
However, aside from the continuation of this and other current trends, there are two other directions of development for corpus linguistics which we consider both desirable and likely to come to pass. The first is that, just as corpus linguistics has become increasingly integrated as a method with other fields of linguistics, it may (and in our opinion, should) be adopted outside linguistics by other disciplines within the humanities and social sciences in particular. Secondly, the triangulation of corpus methods with other research methodologies will be an important further step in enhancing both the rigour of corpus linguistics and its incorporation into all kinds of research, both linguistic and non-linguistic. To put it another way, the way ahead is methodological pluralism. This kind of methodological triangulation is already happening, to some extent, in the case of corpus methods and the methods of experimental psycholinguistics, as we outlined in the previous chapter. But we would argue that it needs to be taken further.

The next three sections outline some particular examples of ways in which methodological triangulation or the linking of corpus linguistics to other academic disciplines (or both) may be achieved. We will look at three specific examples of new developments which have particular promise, in our view. In the next section, we will look at how renewing links between corpus linguistics and computational linguistics may allow new methods in the latter field to assist in corpus analyses. Subsequently, we will very briefly survey some recent work that exemplifies how corpus methods may be of benefit in the humanities and social sciences. Finally, we will look at what we see as the primary challenge for corpus linguistics in the future, that of methodological pluralism, and discuss some recent research which has begun to explore this.

Conspicuously absent from this book has been an extensive review of computational linguistics. Computational linguistics is the field of computer science that looks at how computer systems can be created that work with language in some way. ‘Working with language’ might mean, at the most theoretical level, developing computational models of the language system (see section 8.2.3); or at a very practical level, it could mean developing software that can interact with the user simply by talking to them, rather than by making the user learn to use an interface such as the window-menu-icon systems used in modern desktop software. In either case, software may be developed to analyse language input, as in speech recognition, or syntactic or semantic parsing; or to produce a language output. In cases such as machine translation – automatic conversion of text in one human language to another (see Somers 2003 or Nirenburg et al. 2003 for an overview) – processing of both language input and language output is involved. Another practical problem addressed in computational linguistics is
that of extracting information from a text or texts. As a rough generalisation, we may say that understanding the information within a text is something that humans do extremely well, but slowly; computers, by contrast, do it badly but quickly. If the quality of the information extracted from text by automatic means is improved, then we can benefit in full from the speed of the computer. An example of such an information extraction technique is named entity extraction, where the software attempts to identify all the people, places, organisations and so on mentioned in a text. When techniques like this are applied across a large set of texts, it is often referred to as text mining (see Feldman and Sanger 2007), which is one example of a more general problem called data mining – the identification of patterns and extraction of information across very large datasets. A priority in the field of text mining is currently biomedical text mining (see Cohen 2010) – that is, extracting information from large collections of text (usually academic papers) on biology or medicine. The amount of scientific literature being continually produced in biology and medicine is now far beyond any individual scientist’s ability to keep up with more than a fraction of it; one goal is for text mining software to automatically produce accurate summaries of this vast complex of material for biologists and doctors.

This very short and admittedly incomplete overview of computational linguistics has hopefully served to illustrate that while there are some substantial overlaps with corpus linguistics, it is fundamentally a separate field. Corpus linguistics is ultimately about finding out about the nature and usage of language. While computational linguistics may also be concerned with modelling the nature of language computationally, it is in addition focused on solving technical problems involving language. Computational linguistics as a field converged greatly with corpus linguistics in the 1980s and 1990s (see McEnery and Wilson 2001 for an overview; see also section 4.3) as methods based on corpus data became an essential part of many areas of computational linguistics, and, likewise, advances in corpus annotation enabled by innovations in computational linguistics (such as probabilistic tagging) proved of great value for linguists working with corpora. However, this period of intersection seems to have passed, to some degree, except perhaps in the relatively narrow areas of part-of-speech (POS) tagging, parsing and other forms of tagging such as the annotation of anaphoric relations (see Botley and McEnery 2000). Computational linguistics still makes extensive use of corpora and other sorts of digital ‘language resources’. Indeed, many of the very largest corpora are constructed mainly for use in computational linguistics, for example the English Gigaword corpus, as well much of the other material assembled by organisations such as the Linguistic Data Consortium and the European Language Resources Association. But apart from these shared resources, there is now relatively little interaction between the concerns of the two fields.

We consider this unfortunate, in light of the benefits that earlier cross-fertilisation has had. POS tagging is a near-indispensable tool for linguists’ corpus searches, for instance. Likewise, the development of parallel corpora was in part
driven by the usefulness of such data for machine translation (for example the CRATER corpus, see section 1.7), though parallel and comparable corpora have proven of equal interest for contrastive analysis of languages (see Borin 2001; McEnery and Xiao 2007a, 2007b). However, certain more recent advances in the computational analysis of language have not ‘crossed over’ to corpus linguistics and become commonly used methods in this way, as will be discussed shortly. Computational linguistics is an ‘old friend’ of corpus linguistics, in that they have been and continue to be linked (not least, perhaps, by the somewhat inaccurate perception of those outside the fields of a greater similarity between them than actually exists); but it is a friendship which needs to be renewed and reinvigorated if both sides are to get the most out of the link.

It is, of course, impossible to foresee in full detail all of the possibilities for linguistic analysis that may emerge from a reinvigoration of this connection. However, some predictions may be made based on the nature of research that has continued at the intersection of corpus linguistics and computational linguistics, and by considering potential applications to linguistics of computational techniques that have recently emerged. Much of this work is fundamentally semantic in nature. Research based on semantic tagging, in particular, is notable for having taken place in both disciplines. It has, for instance, been applied as an approach to searching for metaphorical language in corpora (see Koller et al. 2008; Semino et al. 2009). But it has also been applied to problems of information extraction, for example by Sawyer et al. (2002), who use semantic tagging to address the problem of extracting from documents about a piece of software an account of the requirements for that software. We may expect the importance of semantic tagging as a method in corpus linguistics to continue to grow as the implications of the analyses it permits become clearer (Rayson 2008 makes some substantial initial steps in this direction).

Another field of recent interest in computational linguistics, namely sentiment analysis (also know as opinion mining: see Liu 2010 for an overview), takes the computational analysis of meaning far beyond the level of individual words or phrases. The main aim of sentiment analysis may be characterised roughly as the automatic identification of what a writer feels about the topic of the text they are writing (or, alternatively, their opinion of that topic matter). Often, this is directed at the practical task of searching for documents on the web that express a particular opinion, such as blog posts or customer reviews of products in online stores. Since about 2001, this field of computational linguistics has expanded substantially (Pang and Lee 2008: 5–7). Automated sentiment analysis is a non-trivial task for many reasons. For example, whether a text expresses a positive or negative opinion is not necessarily deducible from the number of positively or negatively evaluative words or expressions present in the text. Linguistic phenomena such as pragmatic implicature allow a negative opinion to be conveyed without any straightforward negative expressions such as bad, awful, very poor, I hate X and so on being used (see Pang and Lee 2008: 17–22 for examples).
The key point for our purposes is that sentiment analysis has had little or no impact on the field of corpus linguistics, in spite of some fairly obvious uses for it in discourse analysis and pragmatics. The kind of evaluative language that sentiment analysis looks at has been addressed within (corpus) linguistics (see particularly Hunston and Thompson 2000), but to date this strand of research has remained a somewhat separate undertaking to sentiment analysis. The fault for the lack of integration is not solely on the side of the linguists; Pang and Lee (2008: 13–15) identify applications for sentiment analysis in business and government intelligence-gathering, political science, law and sociology – but make no mention of linguistics. But the potential utility of sentiment analysis for linguists of various kinds is clear. It would be very useful for researchers in pragmatics to be able to search for texts which express a positive or negative subjective opinion. Enabling such searches for a standard corpus such as the BNC will, of course, entail finding solutions to problems beyond those inherent in sentiment analysis itself. For instance, in a general corpus, what is the appropriate unit across which sentiment should be analysed? The datasets of typical interest for sentiment analysis are divided into individual texts that are clearly appropriate units (e.g. individual blog posts or individual product reviews). It is much less obvious that a single text in the BNC, which may be tens of thousands of words in extent, can or should be meaningfully analysed for its ‘sentiment’. But on the other hand, neither is it clear that the sentence or the paragraph (or, in speech, the utterance) would be more suitable units. Nevertheless, the potential value of this approach as an extension to the corpus linguistic toolbox of methods is illustrative of what linguistics has to gain by an active re-engagement with cutting-edge research in computational linguistics.

9.4 The textually mediated world: the humanities and social sciences

The social and cultural world that we as human beings inhabit exists, and is expressed and recorded, to a very large degree by means of language. Socially and politically, we are overwhelmingly concerned with matters (such as right, wrong, friendship, love, justice, freedom, the law, ownership, desires, grudges) that exist only as abstractions within our minds and within the discourse that links our individual minds together. Even non-linguistic cultural experiences such as visual art or music are encountered in the context of a huge quantity of talk and writing about these experiences. Similarly, any knowledge of our culture’s past beyond personal experience exists principally as linguistic (typically textual) communication – and the same is, of course, necessarily true of any aspect of the culture that refers to our future. So it can be said that experience of the human world is largely a textually mediated experience, and, to that extent, human beings live in a textually mediated world.
This is probably the most important reason why many disciplines within the humanities and social sciences are, to a very large degree, concerned with the study of texts. In literary studies, the object of investigation extends from literary texts to critical texts about that literature. The study of religion includes, very prominently, the study of sacred texts and both ancient and modern commentaries on them. History as a discipline is partly based on the physical evidence of objects from the time period under study – which may themselves include text, such as inscriptions on coins or monuments – but also on the study of textual records from the period, or later texts that discuss it. Similar arguments could be made for most other areas of the humanities. All these fields have developed modes of analysing text directed at the particular requirements of the field – as, indeed, have linguists. But for the most part, these analytic techniques are targeted at individual texts. In these subject areas much as in linguistics, however, the information technology revolution of the late twentieth century has meant that very large amounts of text of potential research interest are available in machine-readable form. For example, very large bodies of literary texts are available in a number of online databases. It seems clear that the methods developed in linguistics for handling corpora can be applied to help humanities scholars extend their hand-and-eye techniques of analysis to these much larger bodies of text. In fact, we would argue that any field that is based, primarily or in part, on the study of text can benefit from corpus methods in any research context where the body of text that is of interest expands beyond the point where hand-and-eye methods of analysis can fully encompass its contents.

To give a very basic example, a literary critic who is looking at a single novel may have no pressing need for corpus methods. But if that analyst wishes to look at all the literature of a period, they may well find that corpus methods have something to offer them – not as a replacement for, but rather as an addition to, their existing analytic techniques, just as linguists have found. In writing a study of, for instance, the Victorian novel in English, the researcher’s own experience of critically reading examples of the Victorian novel will necessarily be at the heart of the study. But there is no reason, given the availability of corpus methods, that such a study should be limited to those novels the researcher has read. A corpus of hundreds of Victorian novels could be assembled from text freely available on the web, and used to supplement the core critical analysis. For example, if the researcher wishes to make the claim that the treatment of a certain topic by – say – Dickens is of particular note, then this must obviously be supported by a reading of the Dickens text in question. But it could be reinforced and extended by examining, as a point of comparison, a corpus search for instances of that topic (using a wordform search or a semantic tag search) in the works of all other authors in that period.

Humanities research exploiting corpus tools and resources is a subset of the field of humanities computing, or digital humanities as it is often known nowadays (see McCarty 2005). Digital humanities research includes the development and exploitation of many forms of database, not just corpora. For example, work has
been done to create databases of images (Bailey 2010) and of archaeological objects (Heath 2010). Even when work in the digital humanities is oriented towards text, it does not necessarily treat the text from a primarily corpus-oriented perspective. One central concern is the creation of electronic critical editions of particular documents (see Deegan and Sutherland 2009), where a searchable text may indeed be created, but enabling corpus-type analyses such as concordances, collocations and so on is not the main goal.

However, the sheer size of many of the textual databases that are now available means that corpus methods – and, in particular, tools developed with the aim of analysing the very largest corpora – have a great deal to offer humanities research. We might consider, for instance, the Early English Books Online (EEBO) database, which includes as far as possible everything published in print form in English before 1700. This amounts to many thousands of documents and hundreds of millions of words of text. Indeed, EEBO is larger than any but the very biggest corpora of present-day English. EEBO is obviously an incredibly valuable resource for historians and literary scholars. However, the native search tools of the EEBO interface are oriented to the individual-document approach to text analysis. Searches of the EEBO database for a particular term produce not a set of instances of that term, but rather a list of documents in which that term appears (each of which can then be viewed individually as scanned graphics or as a text file). The utility of such a database could obviously be greatly extended by supplementing (not replacing) these tools with the facilities available in most good modern concordancers (see Chapter 2) such as concordances, collocations, n-grams, keywords, dispersion or distribution analysis, query thinning and so on. Further gains can be achieved by applying corpus annotation techniques to enhance the power of searches. Many corpus annotation procedures have been adapted to historical text, including POS tagging (Rayson et al. 2007) and semantic tagging (Archer et al. 2003). These may require a pre-process to add regularised, spellings as one of the layers of annotation (see Pilz et al. 2008; Baron et al. 2009), in the light of the marked variability in English orthography in earlier centuries.

But if corpus techniques are to be used at all – let alone embraced – across a breadth of humanities subject areas, it is clearly important that the corpora should be made accessible via a user-friendly interface. It is absolutely unrealistic to expect the majority of humanities scholars to get to grips with the theoretical and methodological underpinnings of corpus linguistics as a field – as laid out in the early part of this book – before making use of corpus tools in their analysis. It is similarly unrealistic to expect them to learn computer programming, or the statistics associated with corpus data, or the operation of corpus software tools which are not entirely self-explanatory. As we noted in Chapter 2, these are the directions in which corpus tools are developing in any case, in order to make the methods accessible to all linguists. The necessity of this trend is only that much greater when we consider researchers in other fields.
This level of user-friendliness and accessibility has not yet been achieved. For this reason, so far, studies which are identifiable as applying specifically corpus-based methods to questions in fields other than linguistics have often been done either by, or in collaboration with, specialists in corpus linguistics. Our own research as methodologists has exemplified some work of this kind. McEnery (2005) uses corpus-based analyses to look at historical and sociological, as well as linguistic, aspects of bad language in English; Hardie and McEnery (2009) address text reuse and the expression of bias in early Modern English journalism; Gregory and Hardie (2011) explore procedures by which historical corpora can be mined for geographical information, and the resulting datasets presented in visual form as maps.

However, it is to be hoped that as corpus methods become more accessible to the non-linguist, it will become possible for research like this to be undertaken by humanities scholars without methodological support. The challenge for corpus linguists is to enable this shift to take place.

9.5 The challenge ahead: integrating corpora with new methods in linguistics

As we outlined above, we see methodological pluralism, and methodological triangulation, as critical to the future agenda of corpus linguistics. This argument is, of course, not original to us. We discussed pluralism between corpus methods and experimental psycholinguistics in the previous chapter; but that is not the only context where the case for methodological triangulation has been made. For example, Hollmann and Siewierska (2006) present a rationale for tempering corpus results with other techniques when approaching dialectology. For instance, since most available dialect corpora are small, it is difficult to study relatively infrequent phenomena in them (Hollmann and Siewierska’s example is the ditransitive construction). However, speaker elicitation tasks can be employed to produce additional data which, critically, can be combined together with the corpus data. This is a key point of a methodologically pluralist approach: one type of data does not necessarily trump another, rather different types of data may be used to complement one another – either by confirming some finding, or conflicting with it. In such an approach, multiple types of data may be used to shed light on a single phenomenon. Another example of this kind of approach is provided by Hoffmann and Lehmann (2000), who use questionnaires to test whether or not speakers are conscious of low-frequency collocations (fewer than one hundred instances in the BNC) which nevertheless are highly statistically significant. They are thus able to show that speakers can ‘memorise combinations that are only encountered approximately 5 times over a whole year’ given
an assumption that the BNC is sufficiently representative of speakers’ linguistic experience (Hoffmann and Lehmann 2000: 17, 31). Similarly Arnold et al. (2000), in their corpus-based study of constituent ordering which we reviewed in section 7.3, additionally perform an elicitation experiment which shows that, in addition to the factors identified in their corpus data, a speaker’s personal preference has a role to play in determining the order of noun phrases in a clause. More such work is summarised by Gilquin and Gries (2009). Gilquin (2006) provides an example of triangulation where one method problematises the conclusions we might draw from the results of another; she demonstrates that the corpus frequency of the senses of a word does not, in at least some cases, correspond to the prototypicality of those senses as established by methods of cognitive linguistics. These studies exemplify the kind of methodological triangulation that is already being undertaken with corpus data, and, in fact, a substantial proportion of the literature this book has reviewed could be argued to be (potentially or actually) methodologically plural in this way. In this section, we would like to consider the prospects for new instances of methodological triangulation with other kinds of linguistic research, taking as our example the field of neurolinguistics.

Neurolinguistics may be defined as the study of the language system as it exists in the physiology and neurology of the brain. Core issues in neurolinguistics include: Which parts of the brain are involved in language processing? What are their functions? How do they interact? Neurolinguistics has come a long way since the widely known early work on language in the brain, where the language disorders (aphasia) suffered by people with particular kinds of brain damage — identified via autopsy — were studied in order to deduce which areas of the brain are involved in language. This technique was used in the nineteenth and early twentieth centuries to identify the linguistic functions of key areas of the cerebral cortex, most notably Broca’s area, shown to be important in the production of fluent speech, and Wernicke’s area, shown to have a role in the association of words and meanings. A model of language processing called the Wernicke–Geschwind model, based on this kind of investigation, was prominent for some time (Geschwind 1974). However, there are obvious drawbacks to attempting to model typical language processing in the brain based on disordered language systems and post-mortem study of damaged brains. More recently, neurolinguistic research has been based on brain imaging technology, which allows researchers not only to study language directly in the living brain, but also to use evidence from the study of non-disordered language systems in the development of new models (see Ingram 2007: 59–64). For example, a type of scan called an fMRI (functional magnetic resonance imaging) allows variations in blood flow in the brain to be monitored. This indicates which parts of the brain are ‘working hard’ while a particular task is undertaken, because blood is directed to whatever parts of the brain have the greatest requirements for energy (and thus oxygen) at any given moment. Using an fMRI to observe blood flow during particular tasks in language processing and production allows a profile to be constructed of how different parts of the brain are involved in different types of linguistic thought.
Of course, neurolinguistics does not stand in isolation from other areas of linguistics. For example, those views on language processing in the brain that treat it as a *modular* process (following Fodor; see Fodor 1983) have drawn inspiration ultimately from Chomskyan theory (Ingram 2007: 79). Other links have been drawn between neurology and linguistic theory. For example, Feldman and Narayanan (2004) argue on neurolinguistic grounds that the core semantics of a word is *embodied*. What is meant by this is perhaps most easily understood by a consideration of Feldman and Narayan’s main example, the verb *grasp*. It has been shown that the same groups of cells in the brain are activated both when someone is performing the action of grasping, and when witnessing someone else grasp something. This phenomenon is also observed in non-human primates; the nerve cells in question are referred to as ‘mirror neurons’. What Feldman and Narayanan suggest is that this same neural substrate is the locus of the meaning of the word *grasp*. That is, the process of comprehending *grasp* is a neurological simulation of the action of grasping. This argument is also extended to metaphorical uses of *grasp*, with particular reference to the type of abstract-as-concrete metaphors considered important in Conceptual Metaphor Theory. Metaphorical expressions such as *to grasp an idea* or *to grasp an opportunity* are also understood by neurally simulating the physical action. Feldman and Narayanan link this theory explicitly to the Construction Grammar view of the relationship between form and meaning; they thus create a variant of this theory called Embodied Construction Grammar, where the meaning of constructions is linked specifically to the neural structures underlying the production and perception of physical entities around us (i.e. embodiment as the core aspect of meaning). This theory has important implications for how neurolinguistic evidence should be interpreted; for instance, ‘[o]ne should not expect language to be any more (or less) localized [in the brain] than other perception and action’ (Feldman and Narayanan 2004: 391).

But it is fair to say that, so far, there have not been any significant attempts to link neurolinguistics and corpus linguistics specifically. This is, to a degree, not surprising; at first glance, there seems to be a gaping abyss between the study of text in a corpus on the one hand, and the study of brain scans on the other. But we would argue that in conceptual terms there is actually a great deal of similarity. Our fundamental interest is the *nature of language*. If we assume that language does indeed have the ontological status of a cognitive-neurological system as well as a social phenomenon (as we have noted in Chapter 8, some corpus linguists argue that it does not), the ultimate object of study is the same in both corpus linguistics and neurolinguistics, and in both cases there is a fundamental commitment to empiricism. Furthermore, in both corpus linguistics and neurolinguistics, we study the language system not directly but by observation of epiphenomena – in the case of corpus linguistics we look at the language system’s output on the large scale, in the case of brain imaging we look at either the language system’s blood-flow requirements or some other physiological feature associated with it. So if corpus linguists and neurolinguists have, at least in part, the same object
of study and the same underlying epistemology, methodological triangulation between these fields is, arguably, not only possible but very highly desirable.

Indeed there are certain areas where the findings of corpus linguistics quite evidently require integration with neurolinguistics. Collocation is prominent among these. What does a collocate look like in the brain? What patterns of activity may be observed when collocations are being produced or processed, and how does this differ (if at all) from the activity observable when language functioning by what Sinclair calls the Open-Choice Principle is being processed? Linguistic and psycholinguistic theories about lexicogrammar – whether corpus-oriented, like Hoey’s theory of Lexical Priming, or not – all make certain claims about what is stored in the memory of a speaker. What are the reflexes of these stored entities in terms of the ultimate physiological substrate, that is, interconnections among neurons in the brain? For instance, how does learning a collocation, or being ‘primed’ to use Hoey’s term, act to change the actual structure of the brain? An answer to this and similar questions will inevitably require an integration of corpus evidence and neurological evidence from brain imaging and other techniques. It is not yet clear how such an integration may be undertaken, although it seems clear that psycholinguistics will play an important role in building the link; but it is in our opinion one of the most exciting vistas for future research.

Exactly the same argument could be made for methodological triangulation with psycholinguistics, different forms of functional and cognitive linguistics, and so on. However, as we outlined in the previous chapter, methodological triangulation between corpus linguistics, functional-cognitive linguistics and psycholinguistics is already happening, as is triangulation of corpus linguistic and sociolinguistic methods (see Chapter 5), albeit with much scope for further work; whereas linking corpus findings to neurological findings is clearly a challenge for the future. If this challenge is met, however, the prospect opens up of the development of what we might call a Unified Empirical Linguistics, where evidence of all kinds – textual, psychological and neurological – is as a matter of course used in concert to uncover the fundamental nature of language. It is in the context of such a Unified Empirical Linguistics, we would argue, that corpus linguistics will reach its full potential as a methodology.

9.6 The final word

With this chapter, we conclude our admittedly incomplete survey of corpus linguistics, and its intersection with other types of linguistics and other fields of study. If we have accomplished nothing else, we hope to have given some indication of the very great breadth of corpus linguistics. There are few areas of linguistics where there is no possible role for corpus methods (the most obvious example being Chomskyan theory), and there are increasingly many where corpus methods have become central. At this point, then, it is safe to say
that the place of corpus analysis in linguistics is assured, and that (as we have illustrated in this chapter) the directions in which it seems likely (or desirable) that the field will develop can only lead to yet further expansion of its scope – to new questions, new academic disciplines and new methodological combinations and integrations. Above all, it is this remarkable diversity in the uses of corpus data – and the diversity of viewpoints that these uses may promote – that makes corpus linguistics as a field so utterly refreshing.