GENERAL INTRODUCTION

To compile a Reader for Corpus Linguistics in the Routledge series *Critical Concepts in Linguistics* has been a daring undertaking. From its humble origins about forty years ago, the label *corpus linguistics* has become almost ubiquitous in the academic field of language studies, so common indeed that its meaning is now as fuzzy as that of *freedom* or *democracy*, or for that matter *cognition*. Few linguists today would insist that their findings are not based on real language data. Even many cognitive linguists claim that they work with corpora to explore the workings of the mind's language faculty. But what is commonly denied is that corpus linguistics is a theoretical approach on a level with that of cognitive linguistics, generative linguistics, or structural linguistics, to name only the most prominent frameworks of the past one hundred years.

Is corpus linguistics really a theoretical approach in its own right, or is it, as many linguists never tire of repeating, nothing but a bundle of methods (or at most a methodology) to extract data from corpora which then need to be interpreted in the light of a real theory of language?

Google (accessed on 21 September 2006) lists 1060 occurrences of *corpus linguistics is not*, compared to 9620 occurrences of *corpus linguistics is*. These are some of the citations:

corpus linguistics is not a branch of linguistics, but the route into linguistics
corpus linguistics is not a distinct paradigm in linguistics but a methodology
corpus linguistics is not a linguistic theory but rather a methodology
corpus linguistics is not quite a revolt against an authoritarian ideology, it is nonetheless an argument for greater reliance on evidence
corpus linguistics is not purely observational or descriptive in its goals, but also has theoretical implications

corpus linguistics is a practice, rather than a theory
corpus linguistics is the study of language based on evidence from large collections of computer-readable texts and aided by electronic tools
corpus linguistics is a newly emerging empirical framework that combines a firm commitment to rigorous statistical methods with a linguistically sophisticated perspective on language structure and use
corpus linguistics is a relatively modern term used to refer to a methodology, which is based on examples of "real life" language use
corpus linguistics is a vital and innovative area of research

Some of these citations come from well-established corpus linguists, including Geoffrey Leech and Michael Hoey. In the introduction to their recent edited collection of articles, *System and Corpus: Exploring Connections* (Equinox, 2005), Susan Hunston and Geoffrey Thompson contrast Corpus Linguistics with Systemic Functional Linguistics, and arrive at the conclusion that corpus linguistics "is essentially a method for investigating language", and is "almost perversely 'theory-light' ". Thus it is not only linguists in other camps, it is also the corpus linguists themselves who look at their own field more as a bundle of methods than anything else. Indeed, even at the round table discussion on this question at the 2005 Conference of the American Association of Applied Corpus Linguistics in Ann Arbor, the large majority of participants insisted that corpus linguistics was not a theory.

Do we really need a Reader for corpus linguistics?

Both editors, of course, have been in favour of this project. That is no surprise. And it is also no surprise to anyone who knows us that we both believe that corpus linguistics is indeed much more than an assortment of some computational tools plus some small and large corpora that help us to find things we couldn't find otherwise, things not easily perceptible to the naked eye such as the phenomenon of collocation. That corpus linguistics is much more than this was indeed our firm position at the outset of this time-consuming enterprise into which we had let the publisher goad us. The real surprise was the extent to which the papers we collected vindicated our conviction that corpus linguistics is an approach in its own right. The studies carried out by our authors, often rather modest about the theoretical foundations of their studies, evolve in their cross-fertilisation to a blueprint for a linguistics that breaks with some cherished principles of the past.

The reason Ferdinand de Saussure rejected the mainstream philologists of his time, in full bloom all over Europe, was that he saw in them hunters and gatherers of snippets of linguistic knowledge, not at all tempted to search behind the curtain of language data for an underlying system which would explain those data. This may have been a biased view, for there was actually a keen interest among philologists to establish philology as a real science, in line with the natural sciences. Many "laws" concerning language change were "discovered", and the discovery of many more was predicted. But de Saussure was after something else. His goal was to unveil the secret not of language change but of language itself. Just as biology or physics or chemistry view their object as a system, as something more than a collection of facts, namely as a structure which explains these facts and their relationship with each other, linguistics, he insisted, had to explain its data accordingly. It was the custom of the time, in the late nineteenth and early twentieth century, to compare and equate the social sphere with the natural sphere. Auguste Comte and Emile Durkheim worked hard to establish sociology as a science, and Sigmund Freud's goal was to do the same for psychology.

So de Saussure's revolution can be seen as part of a more general trend of his era.

Is there a language system? What we call language is a reification of certain abstract ideas that people have had in the past, and still have, about verbal communication. The anthropologist Jack Goody suggests that this reification came about only with the invention of writing.

In an oral society, the thought that words are symbols or signs is inconceivable. A sign is always something material, something that has form in addition to having content. Spoken language is a transient phenomenon; it seems immaterial. With the exception of ritualised invocations of the arcane, the idea that an utterance can have a meaning independently of the speaker's intentions would sound odd to the members of an oral community. To understand language as a symbolic system presupposes literacy.

Spoken language, consisting of nothing but sounds and lacking any visual representation, is less "real" than facial expressions or gestures. Yet we do not have a word for communication by gestures. We talk metaphorically about "body language" as if we knew what language is. It was only with the invention of writing that language became a discourse object, a topic one could talk about. In oral societies, language is a social practice that people are mostly unaware of. In an oral society, you do not talk about language. In his book *The Interface Between the Written and the Oral* (1987), Jack Goody maintains that most oral societies do not even have a word for what we call a word. People believe that what is being said are the speaker's intentions. This is what hearers are interested in. Speaking and hearing is simultaneous. People point to things and gesture while they speak. Oral language is full of deictic elements. There are demonstrative, personal, and possessive pronouns, there are locative and temporal adverbials, placing what is said in a situational context. There are, in addition to the (spoken) text itself, intonation, loudness, pitch, accents, gestures, facial expressions and many other features. What do we have to take away from such a situational setting to be left with "language proper"? For all his declared preference for spoken language, de Saussure never made it clear where he would draw the line.

What we call language is an abstraction. What do we mean by calling it a system? There is the tidal system of low and high tide. We can observe its workings in every location where large water masses meet land. We can explain the system in terms of the various laws of nature that are causing it. We can study the lunar influences, the water mass, the tectonic configurations of the coastline and so on. Once we have accounted for all the features involved, we can predict the tides with some accuracy. Is language a system in the same sense? Is there a framework of natural laws, or at least of rules imposed by the language community? Are there "real" entities or other features of language on which these laws or rules can operate? How real are phonemes, morphemes, words or word senses? Features such as these

are constructs that lay people and experts have come up with (and this is a crucial difference between linguistics and sciences like biology: what biologists do does not have to be understood or accepted by non-biologists to make sense, but what linguists do only makes sense if lay people feel that it tells them something about themselves). Yet they are not "things" we can observe. In writing English, we insert a space before and after a word. This makes the word appear to be something real. But many scripts, such as classical Greek or Chinese, do not have these spaces. It is nothing more than custom and convention that makes what we find between the spaces a word. We can easily disagree about what a word is. My version of *Microsoft Word* tells me that there is no word *webpage*. Yet Google lists 75 million hits for *webpage*. The language system is not something "real" in the sense that the tidal system is. We, the people who talk about language, whether lay people or experts, construct these abstract notions. This is indeed what we do whenever we attempt to describe language as a system. But such a system can never be more than a model of our abstractions, not of language itself. If we, as corpus linguists, study language, we do not, like natural scientists, observe and explain facts. Rather we interpret, we make sense of, language as a contingent human artefact. Corpus linguistics is part of the *Geisteswissenschaften*. No doubt there are other, scientific, ways to look at language, for instance neurolinguistics.

If there were a language system, then linguistics has not made much progress towards discovering it over the past century. De Saussure was mostly concerned with meaning, lexical meaning that is. For each language, there is a structure in the vocabulary that shows each word in opposition to all other words. What we have to know about the meaning of a word is not what it refers to in the real world, but what makes it different from all other words. Chomsky's language system deals predominantly with the grammatical structure of sentences. Lexical meaning, to the extent that it features at all in his various versions of linguistic theory, is discussed mostly in terms of word formation. While in the early Chomsky models the universal language faculty determined the grammatical structure decisively (so that all languages are emanations of the same language system), later there seems to be less insistence on concrete language universals, and more on possible variation between languages. In optimality theory, the language faculty provides only the hardware, while each language comes with its own pre-installed software that will be switched on in the language acquisition phase. In the cognitive sciences, the language system is understood as a (modular) part of the mind. The computer provides the model for the mind. Thus we find procedures that translate natural language input into a universal language of thought. The mental representation of a sentence is equivalent to its meaning. Today we find different schools of cognitive linguistics, connected with names like George Lakoff, Dan Sperber and Deirdre Wilson, Ronald Langacker, or Ray Jackendoff. They all have their own ideas about the

language system. The absence of much consensus may well indicate that the various outlines of the language system do not refer to an accessible reality that can be objectively observed.

For if there were a language system, it should be somewhere; and wherever it is, it should be open to inspection. *Langue*-linguistics used to rely largely on introspection, on our competence to determine whether a sentence is correct or not, and if it is not, what it is that makes it incorrect. But how reliable is our introspection? Do we always agree on the correctness of sentences? Does introspection allow us to inspect the language system? De Saussure never came up with a final answer. For him, what is located in the heads of the speakers is a more or less perfect copy of the language system of a given speech community. But which space exactly does this system reside in? Not in the entirety of utterances, not in *la parole*, obviously, because the language system also accounts for utterances that could have been made but were not. If what people have in their heads is only a copy, then where is the master disk?

For Noam Chomsky, the language system is equivalent to our language faculty, and he sometimes even calls it a language organ. This is a specific module in our mind that lets us turn thoughts into language, and language into thoughts. We share it because we are all born with the same hardware. Chomsky was never interested in the contingencies of individual natural languages. For him, linguistics had to be a science. From his perspective it is perfectly reasonable to say, as he often did, that a Martian linguist who came to visit the Earth would be convinced that we all speak the same language, just using different vocabularies.

What actually is regulated by the language system? Obviously it does not predict what has not been said yet, but will be said in the future. According to the theories that posit a language system, this system tells us if an utterance is grammatically acceptable, that is if we can say that it conforms to the laws or rules of the system. Now used we not to say (or: didn't we use[d] to say): "This flat comprises two bedrooms", while now more and more people say: "This flat comprises of two bedrooms"? Did the language system change? Or does the language system admit both constructions, and it is the changing whims of the people to choose one today and another tomorrow? Or are there simultaneously several language systems in operation? To complicate matters further, Eugenio Coseriu, in *Systema, Norma y Habla* (1952), places a third layer between *la langue* and *la parole*: the norm. There is, Coseriu would comment on my example, a more or less eternal system that comprises all the laws determining what can be said in English and what cannot. Within this set of "laws", conventions (man-made rules) determine what may be said and what may not. The usage that finally emerges is *la parole*, that which is actually said. This is not so different from optimality theory. The language faculty is a universal system (and as such, the object of a science of linguistics), but it does not determine the norms

determining a given language, norms which are conformant to the language system.

Langue-linguistics assumes a top-down approach. It presupposes a language system comprising elements, features, categories, relationships between them, and finally rules designed to take care of all "grammatical" utterances. While system linguists have preferred to work, for many decades, with invented example sentences to illustrate the finer points of their models, they now prefer to extract those sentences from corpora, using them as quarries where you pick what you need, and leave the rest untouched. Their language systems are models that claim to translate every natural language sentence into a formal, algorithmic representation. Such models, in principle, can be programmed into computational parsers, devices used in all kinds of natural language technology applications, such as machine translation. But have they ever worked properly? Even though a huge amount of funds was spent over the last fifty years on the development of machine translation systems based on *langue*-linguistics, the results have been far from encouraging.

There is one simple reason for this. Formal languages, like mathematical or logical calculi, are orderly. They clearly distinguish what is a grammatical expression from what is not. They are rigid. Natural languages are the opposite. They are anarchic, lawless, constantly changing, unpredictable. Any attempt to construct a system top-down that will accommodate something that is disorderly and full of idiosyncrasies must fail.

This is where corpus linguistics comes in. Corpus linguistics is bottom-up. It tries to accommodate the full evidence of the corpus. It analyses the evidence with the aim of finding probabilities, trends, patterns, co-occurrences of elements, features or groupings of features. Many of these patterns have been known and accounted for by linguists for a long time. Idioms are an example. We have all learned that there is a phrase *kick the bucket*, and that you cannot replace *kick* by its synonyms *hit* or *strike*, and that you cannot substitute *pail* for *bucket* without turning the idiom into a literal expression. We recognise idioms because they are a cherished part of our linguistic heritage. But idioms are merely the tip of the iceberg when it comes to fixed expressions. There is an infinite number of fixed expressions in our language. There is *cold call*, *friendly fire*, *shotgun marriage*, *social exclusion* and so on. To a large extent they have escaped the attention of *langue*-linguists, because the core element of the language system for de Saussure, for Chomsky, and also for many cognitive linguists, is the single word. The single word seems almost irreplaceable if we want to describe the structure of a sentence. Words somehow seem to be real objects. Compared to them, it appears difficult to determine whether a co-occurrence of two or more words qualifies as a more or less fixed expression. The single word has the advantage of corresponding elegantly to the variables in mathematical or other formal expressions. What is a more or less fixed expression is

always uncertain. How much flexibility, how much variation is allowed? Is *hostile fire* the same as *enemy fire*? What is the relationship of *Iraq's weapons of mass destruction* to *weapons of mass destruction*? Are they two distinct expressions? Compared to them, it seems so much clearer what is a word and what is not.

But words have one big disadvantage when it comes to meaning: they are ambiguous, particularly if they are frequent. Their meaning depends on the context in which they are embedded. There needs to be a context, there need to be one or more collocates, to make their meaning emerge. Once words are embedded in their context, the ambiguity which has puzzled so many linguists simply disappears. The adjective–noun combinations above are, compared to the words they consist of, monosemous. So does it make sense to decompose them into single words? One of the revolutions that corpus linguistics has brought about is to replace the concept of the single word by that of the lexical item, which can consist of one or more words, and which can be described as a unit of meaning. Corpus linguistics also makes us constantly aware that lexical items, be they words or more or less fixed expressions, do not exist as such, but are constructs we posit in our aim to make sense of what has been said.

Corpus linguistics is bottom-up linguistics, is *parole*-linguistics. The starting point is always the corpus, real language data. Whether our analysis will bring order into the anarchy of the discourse is an open question. The statistical analysis of large corpora will find recurrent patterns and other kinds of probabilities. We can measure the statistical significance of co-occurrences. We can observe trends. We can state regularities. But the description of what we find will never yield a language model that is simpler than the complexity of real language data. Do tigers have stripes? All dictionaries say so. Aren't stripes the *differentia specifica* that set tigers apart from leopards or lions? Behold the discourse as represented by Google: there are 255 occurrences of "tigers without stripes" and among them are:

- The existence of black **tigers without stripes** has been reported, but has never been substantiated by specimens or photographs.
- White **tigers without stripes** also exist but are much rarer. These tigers are not albinos as they do not have pink eyes.
- The only wild report to be documented was of a "**tiger without stripes**" at Similpal Tiger Reserve.

Corpus linguistics does away with the certainties of *langue*-linguistics. It questions the view that we can assign senses, or formal descriptions, to single words in isolation. It casts doubt on venerable categories such as parts of speech. It rejects the view that there is an expression in a formal calculus or in a language of thought that is equivalent to the meaning of a sentence. It challenges the expectation that we might eventually come up

with a model of the language system which will allocate its apposite structural reading, that is parse, to every sentence we find in the corpus.

On the other hand, corpus linguistics will tell us how words are actually being used, how they co-occur with other words, and form units of meaning with them. It will tell us that units of meaning have their own unpredictable local grammar, which has so far been largely overlooked. Corpus linguistics demonstrates how registers of all kinds of language varieties differ from each other. It helps us to understand how social reality is constructed in the discourse, how words change their meaning, how spoken language differs from written language, and how children acquire language. Corpus linguistics is constantly searching for better interpretations of language data that make sense to us, the language community, whose communal artefact that language is.

What are regarded as laws and rules in top-down models correspond to generalisations in bottom-up linguistics. Corpus linguistics makes general claims about the discourse, based on the analysis of a suitably selected cross-section of it, that is the corpus. In corpus linguistics, general claims have to do with probabilistic expectations. Unlike claims in *langue*-linguistics, they do not presume the notion of what is grammatical and what is not. They come within the field of grammar, or variation, or language change; and also within the field of lexical meaning, insofar as a text segment occurring in a text can be viewed as an instantiation of a lexical item. If the same lexical item, or any other language phenomenon, recurs in a discourse, then each occurrence can be read as an instantiation of the same type. Each instance can thus be seen as a token of the type constituted by the language phenomenon. It is up to the linguist to define the language phenomena they are interested in. These phenomena do not exist as such, but are constructs designed to answer certain research questions. In corpus linguistics, generalisations normally make claims concerning the co-occurrence of one phenomenon (a word, an element of a group of words, a grammatical category, and so on) with other language phenomena. Thus they show how we perceive language as being patterned. Frequency is an important parameter for detecting recurrent patterns, defined by the co-occurrence of words. It is an essential feature for making general claims about the discourse. However, statistical "significance" is never more than an indicator. It needs to be given an interpretation to become relevant for the language community.

But while over the last forty years generalisations have been the main achievement of corpus linguistics, corpus linguistics can also play an active role in describing each occurrence of a language phenomenon, be it lexical or grammatical, as a unique event. This is particularly relevant if we want to explore the diachronic dimension of a discourse. For what is said today is a reaction to what has been said before, an argument in a simultaneous debate, and an anticipation of what we expect to be said tomorrow. If we look at language from this perspective, we want to make a specific claim.

We want to know what makes a given text segment a unique occurrence, rather than a token of a lexical item type. This will be determined by the unique position that it maintains in the discourse as a whole, embedded in a context that is unique, and referring to a unique set of other texts. In the future, corpus linguistics will develop a methodology to find overt and (more importantly) covert references of one occurrence to similar occurrences in previous and subsequent texts. Unless we can find these intertextual clues that link a given text segment to other texts in the discourse, we will not know what makes it unique.

From the corpus linguistics perspective, the discourse community (not the linguist) is in charge of the language. The discourse community establishes the conventions for what is acceptable and what is not. Linguists are not privileged as 'experts' to pass judgement on what is permissible and what is not. There is no language system that they can point to, no inviolable set of laws and rules that they can invoke. Linguists are part of the discourse community. They have to argue that the generalisations and interpretations they come up with make sense. The discourse community is, in principle, a democratic community. Every member has the right to contribute to the discourse, and to discuss, modify or reject what other members say. Every member can suggest innovations. All the conventions are only provisional and can be re-negotiated at any given time. All regimentation from the outside strangles the creativity of the discourse.

What distinguishes corpus linguistics from *langue*-linguistics are these five principles:

- Corpus linguistics is concerned with meaning, with symbolic content. People are not interested in grammatical constructions; they want to know the meaning of what has been said.
- What sets corpus linguistics apart from cognitive linguistics is that it looks at language from a social, not a psychological perspective. Language is verbal communication between people, is the discourse of what is actually being said (written) and listened to (read).
- Corpus linguistics is diachronic. Whatever is said is a reaction to things that have been said before. We can only fully understand utterances if we know what they refer to. The discourse has, of necessity, a diachronic dimension.
- Corpus linguistics uses frequency to arrive at generalisations. Statistical significance makes us aware of connections that we would not see otherwise. The generalisations that corpus linguistics arrives at are not interpreted as laws or rules, but as plausible ways to group similar things together.
- Corpus linguistics can also make specific claims concerning unique events of language phenomena by showing in which aspects this event differs from all other occurrences of the same type of phenomenon.

Corpus linguistics is *parole*-linguistics. That it rejects the idea of a language system to which language use succumbs does not make it less theoretical than *langue*-linguistics. Nature is always subject to the laws of nature, and the task of the natural sciences is to explore and describe the systems of physics, chemistry and biology. The task of the human sciences, the *Geisteswissenschaften*, is different: it is the interpretation of all that mankind comes up with, ideas, actions and artefacts. Interpretation. The reality in which we find ourselves as members of the human society is not a mirror of the reality out there. It is the reality that has been constructed and is being constructed in the discourse. Language, *parole*, thus is autopoietic: it constantly recreates itself. Therefore the theory of corpus linguistics is hermeneutics, according to Hans-Georg Gadamer not the science but the art of interpretation.

Overview of contents

The purpose of this collection is to encourage, stimulate, and challenge the reader to explore the richness and diversity of ideas and practical applications that have been generated within the field that is increasingly finding its voice under the umbrella of the term Corpus Linguistics. This collection consists of 119 articles and book chapters, divided into twelve sections that cover six volumes and 2300 pages, a testament to the amount of energy that has been devoted to the field in recent years.

The papers themselves were often difficult to track down, as they had been printed in one-off events, for instance in conference proceedings, by the host academic institution, and many such academic publishers have not been able to sustain the published works, which therefore quickly fell out of print.

About 125 authors are represented, as some items are multi-authored, and (partly for the same reason) 18 authors have contributed to more than one item. Although the authors themselves are from many different parts of the world, it is noticeable that the publishers are sited in fewer locations. Some readers may be surprised at the relative paucity of American authors, but this merely reflects the disinterest or disfavour with which Corpus Linguistics is largely regarded in a tradition dominated by generative and cognitive linguists. At one point in the process, we were questioned about the inclusion of so many papers from edited collections rather than refereed academic journals. The reason for this was the absence of academic journals in the field until quite recently (apart from the *International Journal of Corpus Linguistics*), and the wide variety of different fields to which such journals belonged (studies of Text, Discourse, Pragmatics, Applied Linguistics, TESOL, ESP, Second Language studies, and so on), and the papers often involved corpus techniques only at a rather superficial level.

Many commercial publishers charge quite large amounts for the privilege of reprinting their publications. The copyright situation reflects the interests of the publishers rather than the authors. This is not the sign of a healthy situation. Many unpaid hours, and a great amount of effort, went into the negotiation of copyright permissions with publishers. There also seemed to be a reluctance to offer offprints to the authors, let alone copies of the full publication. However, there is a relatively recent and interesting trend for academic authors either to publish their papers on the internet, or to explicitly retain copyright when submitting their work to publishers, which may eventually redress the imbalance.

The earliest article was originally published in 1960, and one is from the 1970s. The 1980s have yielded a handful of articles each year, although there is only one from 1990, and eight from 1999. The main flood of articles comes from 1999 to 2005, which shows the rapid intensification of activities in the field in recent years, after the first "flowering". We have therefore tried to focus on the most relevant publications.

The twelve sections in this collection are:

Part 1 Theoretical aspects of corpus linguistics
Part 2 History of corpus linguistics
Part 3 Corpus composition and compilation
Part 4 Standardisation, alignment, tagging and corpus-related software
Part 5 Lexicography, collocation, idioms and phraseology
Part 6 Terminology
Part 7 Grammar
Part 8 Translation studies, multilingual and parallel corpora
Part 9 Critical discourse analysis / evaluation / stylistics / rhetoric
Part 10 Language history / historical linguistics
Part 11 Language teaching
Part 12 Spoken language / discourse studies

Some readers may be surprised by the fact that we do not have a separate section on Language Variation. However, as will have been evident from the overview of contents, variation is a seam that runs right through all the sections, as the comparison of any two texts or groups of texts is a reflection of language variation. An overview of the contents of each individual section follows.

Part 1 Theoretical aspects of corpus linguistics

As a relatively young discipline, at best 40 to 50 years old, still in the process of establishing its ground rules, tenets, methodologies and practices, it is perhaps not surprising that corpus linguistics is often wrongly accused of being "atheoretical". Using avowedly "bottom-up" approaches (surely in

itself a "theoretical" stance), it might be fairer to say "crypto-theoretical", as it is inevitable that the consensus of generalisations, painstakingly achieved by corpus linguistics after scrutinising vast amounts of data, will only gradually rise up through the numerous levels of abstraction, before arriving at the fully fledged status of "a theory". The views expressed in the Introduction may not coincide completely with those of the contributions in this collection, indeed they may not be shared by any of the contributors. Yet that is surely another indication that the field of corpus linguistics cannot be completely "atheoretical"; only when a discipline has existed for some time can there be a consensus about the ideas it espouses; until then, it is unsurprising or even de rigueur that opinions vary, or are at odds, or even directly contradict each other.

Chafe sees corpus linguistics as proceeding from understanding language to understanding mind. He constructs a matrix of behavioural and introspective procedures, acted on by artificial manipulation and natural observation techniques. **J. Aarts** asserts the existence of corpus linguistics (referring to Fillmore, who described theoretical linguistics as "armchair linguistics"), and contrasts intuition with corpus data, and corpus-based with corpus-driven approaches, using spoken data as a case study. **Tognini-Bonelli** explores the corpus-driven approach in greater depth, calling it a qualitative (*sic*) revolution. No data are excluded, and descriptive categories are based on recurrent patterns and frequency distribution. There is no discontinuity between text, genre, variety and contexts of situation and culture. The importance of computer technology, and the unity of form and meaning, are highlighted.

De Beaugrande, like others in this collection, criticises Chomsky and asserts that corpora reveal not the disorder of language use, but its different modes of order. He takes 20 arguments against corpus linguistics and refutes them systematically, indicating instead how corpus linguistics bridges many of the gaps between theory and practice. **Knowles** says that accusing corpus linguistics of being "atheoretical" is a defensive position, and explores the nature of its threat. Theory is a stance towards the organisation of data. Corpus is a distinct stance within the accepted relational model. Chomsky's theory is merely a different stance. Mainstream linguistics consistently backgrounded language texts. The corpus approach potentially undermines conventional assumptions about the nature of linguistic theory. **Teubert** emphasises the role of semantics in corpus linguistics, contrasts oral and literate societies, and dissociates language from both the "real world" and the "minds of the speech community". He suggests a hermeneutical approach, and predicts that the next focus of corpus linguists will be on the diachronic continuity and uniqueness of meaning of a lexical item within the history of the discourse.

Lindquist and Levin warn of the danger of being perceived as "counting for counting's sake", foreground the research question, and see corpus as a

means to an end, and therefore the choice of an appropriate corpus as the key. **B. Aarts** examines the tension between empirical, corpus-based linguists and theoretical linguists. He quotes Chomsky: ". . . arrangement of data isn't going to get you anywhere", and suggests that corpus linguists should focus more on "the qualitative data that corpora can furnish". **Sinclair** focuses on meaning, posits the lexical item as the principal unit of meaning, and asserts that it is monosemous. However, the combination of such items is unpredictable, and meaning is imprecise and provisional, hence the lexicon can never fully account for all the meanings it can generate. Meaning is also generated at the sentence-level, by increments. Language is therefore processed simultaneously at both levels. **Fillmore** concedes the benefits of corpus data to an armchair (theoretical) linguist. Corpora are at the same time linguistically inadequate (non-comprehensive) and uniquely revelatory of linguistic observations that could not be obtained by any other method.

Part 2 History of corpus linguistics

The brevity of this section is merely a testament to the fact that the history of corpus linguistics is relatively short, partly associated with the recent advent of computer technology.

Sinclair and Jones relate that the study of lexis, and in particular collocation, could not begin until this technology arrived. The experiments still centred around sentence invention, elicitation, and word association, but the main finding was that intuition does not – or perhaps cannot – produce the typicality found in naturally occurring texts. This early research involved taking "arbitrary" decisions, in the absence of established methods, but strongly indicated a form of pattterning (collocation) that could not be subsumed under syntax or semantics. **Teubert**'s interview with **Sinclair** (40 years later) emphasises that the data involved were of spoken, not written, text, and the focus was on lexis (rather than grammar, the main linguistic interest of the time). Theories of speech and lexis were almost non-existent. The research undermined several traditional beliefs: that grammatical words did not have collocations, that all forms of a lemma shared the same collocates. Discoveries were made: that position of collocates was not significant. But technology and software were still inadequate, and even the researchers found it difficult to abandon mainstream beliefs (e.g. the word as a unit of meaning). Statistical significance continues to be a problem, as does homography. Collocations are neither rule-based nor invariant. Distinctions are often simultaneously grammatical and lexical, not either/or. "If there is no choice, there is no meaning." Idioms are interesting, but (because?) they are rare. "We need to fit the forms to the meanings, not the other way round."

Francis broadens the contemporary dictionary definition of the purpose of a corpus to "linguistic" rather than "grammatical" analysis, and reflects on the contemporary (still current?) focus of linguistics on "competence"

rather than "performance". He refers to several users of the corpus as studying matters which "can and in some cases must" be studied from performance data. As suggested in the title, the focus is on problems; problems of population and sample, inequality of text reception, the incapacity of corpora to include *all possible* (grammatical) utterances (the holy grail of some grammarians), the role of the computer, homography and polysemy, and lemmatisation.

Quirk reflects on the inadequacy of data for the description of English grammar, and therefore of the description itself. In the context of English pedagogy, this allows teachers a great deal of latitude, a "rebuke and challenge to linguists". Contrast the wealth of information in dictionaries – but even The New English Dictionary used to "concoct sentences and phrases" to illustrate "closed" category (grammatical) words. Account had to be taken of *all* the data in the Survey of English Usage. The word is "fully institutionalised", hence the preferred minimal unit of research. Bottom-up procedures are outlined. The inclusion of meaning as a focus for linguistics is gradually accepted. Mention is made of factors affecting interpretation, and their dependences. Educated natural usage is supplemented by educated reactions to such usage (both "believed" and "perceived" usage).

Leech suggests that literary scholars and historical linguists assume the validity of corpora. Only synchronic linguists have claimed that their intuitions are a sufficient data source: "the primacy of intuition remains an orthodoxy". The corpus is an important – but not unique – source, but is essential for some types of research, where intuition is totally inadequate. Data are the basis of most other sciences, yet Chomsky has succeeded in reversing this notion. Labov (and sociolinguists in general) oppose him. Leech sees intuition as suited for interpreting linguistic data, but not for generating it. Intuition can generate data not found in corpora, but corpora can equally reveal data not retrieved by intuition. Intuition is unsuitable for sociolinguistics, pragmatics, and text linguistics. Corpora reveal gradiences between categories that intuition regards as absolute. Intuition is as skewed as corpus data, sometimes focusing on prototypical, sometimes on marginal instances. Corpora reveal facts that are below the level of consciousness accessible to intuition (exemplified by case studies of near-synonyms). Case studies of near-synonyms. De Saussure's *langue* (language as a shared social phenomenon) is more useful than Chomsky's competence (individual psychological phenomenon). Elicitation from native-speaker informants can supplement intuition and corpus. Lexicographers and literary scholars have long known the value of corpus. Grammarians need to recognise its value as well.

Léon dates the rise of the term "Corpus Linguistics" to the 1990s, and cites the rapid increase in publications and the emergence of a journal (*IJCL*) as evidence. However, she suggests that its proponents have distorted

history in order to legitimise it as a discipline. "Any linguist is a potential user of corpora." She says that some corpus linguists claim empirical, statistical connections with the 1950s, displaced for 25 years by "rationalism" (Chomsky) and awaiting the liberation provided by computers. The 1970s saw a shift from knowledge-based and rule-based methods to probability-based ones. The Brown corpus is accepted as the first (ignoring the Survey of English usage; the Trésor de la Langue Française for French; and the Rand corpus). She says that the Brown corpus claims to be "for grammar studies", but Kucera and Francis's book focuses entirely on statistical studies. There was no "discontinuity" in corpus studies, and Chomsky was not responsible for it. Quirk and Svartvick support performance and acceptability over Chomsky's competence and grammaticality. Chomsky said that the use of corpora reduces linguistic description to a mere list without any explanatory hypothesis. Leon suggests that "Two retrospective constructions have been forged . . . a theoretical anti-precursor . . . i.e. Chomsky: and a technical precursor, in fact a product, the Brown corpus. Why does corpus linguistics need to be an autonomous discipline? Chomsky was not attacking corpora but the use of Markov models for higher-level linguistic units.

Part 3 Corpus composition and compilation

Advances in computer technology, and the increasing availability of data already in electronic form, have greatly eased some aspects of the process of compilation. However, technology brings its own share of problems and challenges: some electronic formats are difficult to convert for processing, separating non-text-data from text can be complicated, and so on. The external and internal selection criteria for corpus composition have now been discussed at some length, but remain problematic, inconsistent, or conflicting. Potential corpus sizes have increased as a result of fewer technological constraints, but a lot of interest is also being shown in smaller, well-defined datasets. Automatic corpus creation from the web has increased recently.

Leech contrasts the "closed" corpora for dead languages, and the "practically limited" corpora for living languages. He counters several Chomskyan arguments against the use of corpora, concedes the limitations of corpus approaches, outlines the impact of computational technology, then surveys several corpus projects (Brown, LOB, London-Lund, Survey of English Usage, FLOB and FROWN, British National Corpus, ICAME, LDC, COBUILD's Bank of English, CHILDES), discusses problems of availability, annotation, and so on, and looks at specific application areas (lexicography, language teaching, translation, and speech processing).

Erjavec describes the creation of a Slovene-English parallel corpus, emphasises the use of open standards and publicly available tools and resources, outlines the processes (normalisation, tokenisation, segmentation,

alignment, POS tagging and lemmatisation) before discussing two output applications (web concordancer and bilingual lexica). He also mentions the cyclic nature of corpus development.

Burnard looks back at the BNC, positioning it in its computational context (WordPerfect vs WinWord, PCs with 386Hz processors and 50Mb hard disks, Unix networks, etc.) and its corpus development context (three trends: Brown/LOB/ICAME, Birmingham/COBUILD, and computational linguistics), with lexicography as the main application. The converging interests of Humanities and Computer Science were stimulated by European Union funding for Language Engineering. The BNC set many benchmarks for industrial standardisation in corpus development. He mentions the unexpectedly enthusiastic uptake of BNC by applied linguists (rather than computational linguists), its widespread use in language teaching and learning, by single users rather than in networks. The rapid advances in computer technology, and distribution of audio files were overlooked. The BNC is a source for many specialised subcorpora, but lacks a "monitor corpus" function to track language change.

Johansson and Hofland describe the development of a parallel Norwegian-English corpus, of general linguistic (language universals and language typology) and language-specific value. The corpus consists of original texts in both languages and their translations in the other language. Alignment was at the sentence level, using sentence length and anchor words. The primary interest was in translation and translationese, and genre. **Reppen and Ide** report on the development and first release of the American National Corpus, which adopts many of the principles of the BNC, firms up the standards for text encoding and annotation formats, and will form a basis for comparisons of British and American English. **Mair** shows how parallel corpora of different vintages can be used to investigate language change, on the basis of analyses of the Brown and LOB corpora (1961 texts) and the Frown and FLOB corpora (1991–2 texts). The analyses focus on perceived grammatical changes, but the conclusion is that the changes are largely a result of the colloquialisation of written norms, changes in selectional preferences rather than in the structures themselves.

Greenbaum charts the development of the International Corpus of English, a project responsible for the compiling of corpora of 1 million words in 13 countries in which English is the first or second national language. The corpus design mimics the Brown and LOB corpora, 500 texts (by educated adults) of 2000 words, originating in 1990–3. A policy was established for transcription of speech. Concordancing software will be provided, and POS tagging and parsing will be implemented. The data will primarily benefit sociolinguists comparing national varieties. **Atkins, Clear and Ostler** discuss corpus design criteria in general, without reference to a specific corpus. Terminology for data collections and units of text are defined, stages in corpus building are outlined, and the basic software tools are listed and described. Issues

such as copyright, sampling principles, corpus typology, text typology, and mark-up (annotation) are discussed and standards suggested. The various uses of corpora and the interests of users are also considered.

Biber addresses the core issue of representativeness in corpus design, which underlies all assertions that a corpus is a valid basis on which to make generalisations about a language or a subset of a language. He considers stratified and proportional sampling techniques, and sampling within texts. He argues that theoretical research is required prior to corpus design, to identify the situational parameters that distinguish between the texts of a speech community; complemented by empirical investigation (univariate and multivariate) of the linguistic variations (involving registers and text types) in a pilot corpus; which should then form part of a cyclical process in corpus development.

Granger considers the development of corpora of learner English as a resource for research into second language acquisition, contrasting corpus data with the traditionally used introspective and elicited data, and linking it with error analysis data. Native-speaker corpora may approximately represent the target model, but learner corpora reveal the problems and patterns of learner language. She also looks at corpus design issues and types of analyses, and suggests that learner corpora will throw light on some unresolved areas such as the role of transfer, as well as enhancing pedagogic tools and classroom practices. **Feng** asserts the primacy of "real language data", focuses on corpus linguistics in Chinese, but subscribes to its lack of "commonly accepted and fully developed theory".

Part 4 Standardisation, aligmnent, tagging and corpus related software

This section looks at the addition of annotated information to raw language corpora, involving manual, semi-automated, and automated processes (and hence is probably the most computational or technically oriented section in this Reader). For example, annotating each word in a corpus with a wordclass (or part-of-speech, POS) tag has become a commonplace procedure. However, every process is bound to generate its own set of errors and inconsistencies. And as different systems develop for each form of annotation, the need arises (in terms of the reusability of the corpus) for some standardisation or conversion techniques.

Van Halteren proposes a method for detecting inconsistencies in manually POS-tagged text. An automatic tagger is generated from the corpus, then applied to the manually tagged corpus, and non-matching items are flagged. However, van Halteren concedes that this technique may not be so easy to implement for other linguistic features, such as word senses and syntax. For these, he suggests using several automated systems, and inspecting items where more than one system disagrees with the manual annotation.

Kilgarriff suggests that corpus linguistics lacks a corpus taxonomy and typology, and lacks quantitative strategies for describing and comparing corpora. For descriptions, he offers internal (linguistic) criteria, using word and n-gram frequencies and a ranks test. For comparisons, he evaluates various possible corpus similarity measures, and proposes a Chi-squared one as the most suitable. He concedes that work is needed to embed his suggestions within an appropriate wider mathematical model, and to make them scale-independent, allowing comparisons of small and large corpora.

Véronis surveys the processing of parallel corpora (texts and their translations). He considers the techniques for aligning texts at different levels (sentence, clause, word) and evaluates them. He also looks at the use of parallel texts in various fields (translation, lexicography, information retrieval), and the availability of corpus resources. His references to the rapid increase of multilingual information on the Web and of global markets have been more than vindicated even in the few years since his publication.

Church was an early researcher in this field, but is already aware of the problems of aligning parallel texts at sentence level (especially from the noise created by OCR output and unknown markup conventions), hence he proposes a method for aligning at the character level, using a cognate approach. This avoids the need to identify sentence and paragraph boundaries, and is claimed to work quite satisfactorily.

Scott investigates the analysis of key words at the text level, in the context of two PCs: personal computers and political correctness. Words that are key in many texts are termed "key key words", and their associates (words that are key in the same texts as the key key word) are shown to reveal aspects of text schemata and stereotyping in relation to socially important concepts. Using newspaper texts, Scott shows that clumps (produced by a crude procedure requiring refinement) of associates characterise stereotypical attitudes towards these concepts. The procedure may have implications for text retrieval, language pedagogy, critical text analysis and literary criticism.

Fang considers the problem that several automatic grammatical tagging systems have developed independently, and therefore in order to work with several existing corpus resources, a cross-tagset mapping procedure is needed. This will not only make more resources usable to more people, but also enhance higher-level grammatical processing such as parsing.

Sperberg-McQueen *et al.* look at formal markup (SGML) systems. Markup signals the occurrence of distributed (logically non-countable, e.g. use of italic font) and non-distributed (standard text structures such as paragraphs) features. Markup allows users to make inferences about the marked-up passage of text, and various approaches are outlined in relation to the problem of interpreting the meaning of the markup at specific locations in the text. Markup may be inserted by authors, or by transcribers creating electronic versions of pre-existing texts. The first inference is that the

contents of the marked up passage shares some property. Different elements in markup are discussed and the permitted ("licensed") inferences are discussed and collated into a framework. The possibility of discovering a common vocabulary (semantic primitives) for the elements of different markup systems could result in easier automatic conversion of texts across markup systems.

Bird and Liberman survey a wide variety of annotation formats in use (from temporal markers in speech data to textual annotations such as phonetics, POS, named entities, co-reference, and other discourse features), and focus on a logical framework rather than electronic file formats. The survey includes the goals of annotation, and cost-benefit analyses, and results in a framework containing a set of ideal formal criteria for all annotation systems, comprising the common conceptual core, the annotation graph.

Nagao describes existing machine translation systems as "inherently inconsistent" and proposes a model for MT based on the use of analogical thinking, using Japanese and English as exemplars. Humans start by memorising some initial examples, then noticing similarities and differences in a variety of examples, guessing and making inferences. MT systems should do the same. To speed up the process, systems are given a lot of the information (redundant expressions, sentence structure, word and phrase dictionary with grammar, meaning and verb frames, word thesaurus, global sentential structure and local phrase structure) in initial system construction. The learning by analogy is required only at augmentation stage, when increasing the range and number of example sentences and improving the thesaurus.

Sinclair calls into question the whole process of annotation, contrasting "corpus-driven" linguists who prefer un-annotated data (having doubts about the validity of "intuitive" annotations) with corpus-based ones who regard annotation as indispensable and, faced with a straight choice, might even prefer the annotation to the text. The former are characterised as cultivating "degeneralization" (deferring intuitive responses) to allow some degree of objective independence from the data. Annotations are usually based on "pre-corpus" linguistic models, and often require substantial manual manipulation in order to "fit the corpus data", and are therefore seen as acceptable for applications requiring quick results, where rough-and-ready methods are sanctioned, provided that the limitations are recognised and certain safeguards are respected.

Clear looks closely at computational methods for investigating collocation, highlighting the fact that the stereotyping of word combinations is a pervasive feature of language. Statistical methods are discussed and collocations are characterised in terms of their frequency, idiomaticity and positional variation. Collocation occupies an ill-defined area of linguistic patterning that is neither clearly syntactic nor clearly semantic, and neither lexicographers nor theoretical linguists have been able to establish appropriate parameters for collocation, but computational methods may offer potential solutions.

Oakes and Lewandowska-Tomaszczyk propose using bilingual and trilingual alignment and concordancing as aids to human translation, extending Gale and Church's alignment method to the trilingual context and using Scott's Wordsmith Tools to create and display the concordances. The results for trilingual alignment (90%) were lower than for bilingual (98%). Individual words and phrases were searched in English, Polish and French versions of Plato's *Republic* and proved potentially useful to human translators and language analysts.

Gale *et al.* consider the problem of polysemy and automated word-sense disambiguation, and suggest that if a polysemous word occurs several times in the same text/discourse, it is highly likely to be used with the same sense in most cases (98%): hence, one sense per discourse. Using discourse-constraints will improve the performance of word-sense disambiguation algorithms, and help to evaluate those not using such constraints. Sense-tagging of polysemous words in a corpus becomes more possible, especially in view of another finding: most words are not highly polysemous as previously thought. Indeed quite the opposite is the case: most words have only one sense. So the problem of word-sense disambiguation may not be as insoluble as perceived hitherto.

Part 5 Lexicography, collocation, idioms and phraseology

This section is naturally substantial, as the study of lexis, collocation, idioms and phraseology is exactly a mirror of the 'bottom-up' approach that characterises corpus linguistics in general, and also reflects the fact that lexicography was the first field to use corpora extensively in updating its methodology. However, corpus linguistics rather than corpus lexicography remains the focus, hence the articles relate to linguistic features rather than specific dictionary projects.

Sinclair asserts that word and sentence are the two "primitive" units in the language form of written texts; that sentence is the unit that aligns grammar and discourse, and word is the unit that aligns grammar and vocabulary; and that word is the starting point of the description of meaning in language. Other units are discussed (e.g. morphemes) as well as models of arrangement vs process. Traditional linguistics focused on grammar rather than lexis. Lexicography has always had to deal with multi-word units of meaning such as compounds, phrasal verbs, phrases and idioms, and especially collocations, which unfortunately do not fulfil formal criteria. Various multi-word sequences are used as examples in the search for units of meaning.

Louw examines semantic prosody ("a consistent aura of meaning with which a form is imbued by its collocates") and argues its diagnostic potential for irony and insincerity of various kinds. This seems to be a linguistic feature that introspection alone is almost totally unable to retrieve, and is only observable with the help of extensive corpus data. The potential is seen as lying in

the "composition and grading of suasive language" and radical stylistics. Semantic prosodies co-occur to reinforce each other, and other features (e.g. metaphor) may also act as signals. The extent of pervasion of the language by semantic prosodies is as yet uncertain, as they are not easy to detect. However, the impact on the study of suasive genres such as propaganda and advertising, and in revealing false, insincere, or indefensible claims, are obvious.

Berber Sardinha presents a contrastive study of the semantic prosody (seen as connotation) of equivalent items in English and Portuguese, points out similarities and differences, and reflects on the inadequacies of dictionaries in recording such information. Semantic prosody is created by multi-word associations, and needs to be recognised by human translators (and machine translation systems), as they seem to be distributed unevenly and even unsystematically across near-equivalent terms in different languages, thus raising the possibility of creating a connotation dictionary.

Deignan explores the semantic relations of lexemes in their literal and metaphorical uses in the domain of temperature/emotion, and concludes that although there is a systematic maintenance of semantic relations at the more general/abstract level and with less frequent lexemes, deviations and inconsistencies occur when examining specific uses and patterns, especially with higher frequency lexemes. Literal-metaphorical mapping of semantic relations is far more complex than anticipated.

Marcinkevičienė continues the theme of searching for appropriate units of meaning, confirming that word is not a satisfactory unit. The terminological tendency, the open-choice principle, creates an extended term bank with fixed meanings and clear differentiation; the phraseological tendency, the idiom principle, creates an "empty lexicon", based largely on context-dependent usage. The latter requires more attention and effort, and "patterns of usage" are posited as a focus of interest (involving both syntactic patterns and collocations) between corpus concordances and elaborated dictionary entries. Such patterns of usage would need to be linked to meanings via semantic preferences and prosodies.

Mason argues that each word form has a measurable impact on its textual environment, and posits the notion of "lexical gravity" for this influence. The size of the environment can be established by looking at the variability of words within it. After investigating the various parameters involved in collocation (size of corpus, corpus selection criteria, environment/span, cut-off/threshold, node/collocate preprocessing – lemmatisation, POS tagging, case sensitivity, etc., significance measure, reference frequencies), the focus is on how to establish the environment/span.

Previous studies assume a span of four words either side of the node. Variability is obviously zero at the node and increases steeply away from the node. A detailed examination is conducted using type-token ratios (TTR) for each "slot" away from node, indicating both lexical gravity (degree of

selectional restriction/variability for co-occurring words) and optimal span. Inflections affect gravity, as does wordclass. The effect of word sense requires tools which do not yet exist. Multi-word units and other languages need to be investigated, as well as alternatives to TTR.

Calzolari looks at corpus-based lexicon building (an expanding field as a result of more available corpora), the impact of corpus-based print dictionaries, and better software tools. Examples are drawn from large multilingual European projects requiring harmonised lexicons, with a focus on acquiring lexical information from corpora to enhance existing lexicons. The corpus and the lexicon are seen as the platform on which human language technologies will need to be based in future, and community/industrial needs in a global market are highlighted.

Hanks discusses whether word meanings can be identified empirically, and argues that although corpus analysis has suggested that norms of usage (which he contrasts with "exploitations" such as metaphor) can be established for each word, the criteria for distinguishing norms from exploitations have yet to be defined. The problem of changing norms over time is also considered. This diachronic study therefore looks at the differences between norms in a historical corpus and a modern corpus.

Moon investigates the distribution of idioms across text types, building on earlier observations that idioms generally occur very infrequently, occur more in journalism than in other written texts, and might occur more frequently in spoken data; and that idioms are highly variable and often manipulated. Limitations of the study included the small number of idioms examined, and variables such as syntactic form, degree of informality, meaning and evaluative orientation which could not be fully explored or calibrated. Idioms were found to occur more in contexts of constructed interactivity and informality, in suasive texts; and individual idioms had individual distribution patterns, reflecting meaning and pragmatics.

Stubbs examines the idea that words have distinctive semantic profiles or prosodies, and that the strength of association between words can be measured in quantitative terms. The focus is on the various quantitative methods, the collocational sets established, and the theoretical basis which might underlie them, especially in highlighting the relationship between system and use. The case study suggests that prosodies may arise from the influence of *parole* on *langue*. The results reveal a relationship between lemmas and semantic categories hitherto unrecorded in dictionaries or grammars.

Ooi explores collocations in Singaporean and Malaysian English in contrast to those in British, American and Australian English, to see whether they reflect the differences of realities and values in these areas of the world, and could create a better understanding of intranational and international communication and a better global English dictionary. Although some cultural differences are suggested by the data (even between Singaporean and Malaysian English), conclusions are necessarily preliminary, as some of

the realities and values are no doubt encoded in the other languages used in these areas, and they in turn may exert different influences on the local English varieties.

Barnbrook tackles the sublanguage of dictionary definitions, which try to express unknown features of language in terms familiar to the user, and are of crucial significance to the effectiveness of the dictionary in general. In particular, he examines the relationship between definitions in specific dictionaries and the needs and expectations of their users. The learner dictionary's definitions (as opposed to those in the native-speaker dictionaries) are found to be less complex and more consistent, using a more limited range of lexis, and with the more commonly used words forming a larger proportion of the definition vocabulary.

Altenberg reports on the phraseology of spoken English, with evidence of recurrent word-combinations of 3+ words occurring 10+ times (arbitrary cut-offs, as neither length nor frequency are criteria of phraseological status) in the London-Lund corpus. Short and infrequent phrases are therefore automatically excluded. The results show the pervasiveness and variability of conventionalised language in speech, from whole utterances at discourse level to multi-word units acting as single words. Few examples are semantically or grammatically "frozen", illustrating the overlap between lexicon and grammar. Higher level units tend to have pragmatic functions, lower level units have propositional (lexical, grammatical) functions. At clause level, sequences of clause elements appear in recurrent clusters reflecting routinised ways of presenting information in speech. The "fuzzy" nature of phrases is emphasised.

Hoey asserts that corpus linguistics has not attended much to text-linguistic issues; that lexical choice has a major effect on cohesion, theme choice and paragraph division; that some lexis is biased towards certain textual functions; and that lexical choices interlock to create colligational prosody. Within his model of "lexical priming", he outlines a new theoretical relationship between lexis and text-linguistics, with textual colligation at its centre. The choice of lexical item co-selects its primings, which may be positive or negative with respect to cohesion, semantic relations in text, theme, and textual divisions.

Atkins *et al.* seek to identify the essential components of a word's context (which should therefore be recorded in any dictionary database), from the theoretical perspective of "frame semantics". The categories of lexicographically relevant information are defined. Analysis includes the semantic content of the word, identification of its semantic neighbours and the differences between them, and grammatical constructions in which it takes part. The obligatoriness or optionality of these elements is also significant. The focus is on "frame-evoking words". The object of the study is to assist lexical analysts looking at vast amounts of corpus data to focus on significant information categories, and also to facilitate in part the automation of the task.

Part 6 Terminology

This section is short, because terminology is really a sub-field of corpus linguistics, and overlaps substantially with several other neighbouring disciplines, such as translation and computational linguistics. **Bowker** describes the emergence of corpus-based terminography, the process of retrieving terminologically interesting information from corpora, and its specific needs in relation to corpus design, methodology, and tools. Terminographic corpora need to consist of LSP texts rather than general texts, and computational procedures are seen as reducing some of the more tedious and labour-intensive tasks, and facilitating rather than replacing the careful manual analysis still required.

Williams says that representativity in LSP corpora is usually measured by external selection criteria. To overcome Subjectivity, he suggests that corpus-internal selection procedures should be adopted using lexical criteria. He uses restricted collocational networks to group texts within special language corpora, and finds that audience is a major factor in strong and weak prototypical groupings in both theme and domain-specific corpora. Domain-specific journal texts are more central than theme-specific conference proceedings. Using Clear's terms, "clues" and "antis", he finds that antis are not necessary, but dispersed lexical and non-lexical items grouped as complex nodes can be essential discriminatory elements. He warns that we must first analyse the texts carefully to establish the lexico-grammatical norms, often involving less obvious semi-lexical items. Terms alone are not sufficient to define disciplines and Sub-domains. Rather, disciplines may make more or less use of certain terms, thus revealing their particular focus. Frequency of lexical subsets such as gene names, chemical formulae, temperatures, or specific activities may typify specific text groups. Unlike general language corpora, there can be no one-off generic diagnostic for categorisation of texts in LSP corpora. The traditional equating of term-concept-domain does not work.

Chodkiewicz *et al.* argue that computer-assisted term extraction requires further lexicographic treatment by humans, using linguistic and subject specialist knowledge, based on a study of a human rights glossary for translators. The corpus term extractor identified candidate terms in French, but a legal expert had to discard irrelevant sequences and pair the English equivalent in aligned text. This process yielded many multiple correspondences in both languages, some of which could be resolved purely by linguistic means, but others required further expert knowledge. However, automation does have specific benefits: frequencies are very useful, as is the ability to access all texts containing a particular term; the system prioritises multi-word units, which usually have fewer multiple equivalents; analysis of verb groups significantly reduces multiple equivalences. **Pearson** reports the benefits of using a corpus of specialised texts for terminographic work, and presents criteria for selection of appropriate texts, and also for identifying definition elements

embedded in the texts, using specific grammatical structural features, or via concordances for the term (which, however, requires more manual effort).

Part 7 Grammar

Corpus linguistics generally focuses on local grammar and valency grammar. It is also frequently used to critique traditional grammars. **Römer** examines the occurrence of progressive forms in native-speaker spoken English corpora and in German EFL textbooks. Significant discrepancies were found, and surprisingly greater variation between the German textbooks. These differences, added to the deficiency of traditional grammatical descriptions, may account for many learner problems in this area, and may be substantially reduced by using native-speaker corpora to amend the textbooks.

Mahlberg adopts a text-linguistic approach to look at aspects of the support function of general nouns (such as "man, move, thing"): giving emphasis, adding information in passing, and providing an introduction. The text-linguistic view requires integration of structural description (the "pattern grammar" approach) and functional interpretation. Corpus linguistics needs to move from the lexico-grammar level focus to a text-level focus.

Kennedy asserts the semantic complexity of the most frequent (structural) words of English. Native-speaker intuitions and pedagogic works tend to focus on locative uses and areas of overlap. This study uses statistical information to disentangle the detailed descriptions of systemic possibility given in dictionaries and grammars, in order to counter the potential arbitrariness and unreliability of intuitive judgements and produce more pedagogically sound and useful outlines. The items examined in detail show that there are significant differences in the collocations and functions actually observed in usage.

Mindt asks when we can invest confidence in a "grammatical rule", what the status of exceptions and errors are in language description, and what inferences we can draw for language change from the structure of a grammatical rule. Of the many possible realisations, only three make up the core of an individual grammatical phenomenon. The core realisations are not evenly distributed. The core makes up approximately 95 per cent of all cases. The approximately 5 per cent remaining cases ("exceptions") are usually errors, obsolescent patterns, or emerging patterns. A clear distinction can be made between grammatical rules and the behaviour of lexical elements.

Conrad brings together two recent historical developments: the renewed interest in grammar teaching, and the availability of new corpus-based grammatical descriptions. She suggests that the latter will prompt three changes: a shift from monolithic grammars to register-specific grammars, the integration of grammar and vocabulary teaching, and a shift of focus from structural accuracy to appropriate contexts for alternative constructions.

Corpus studies indicate the variation in frequency and use of constructions, and the complexity of grammatical choices (not simply decisions about formal accuracy). The variations in native-speaker use are highly systematic in relation to social and linguistic context, and are very useful for teaching and learning English.

Hoey moves from description to theory to practice. He suggests that the traditional separation of grammar and lexis in teaching materials is unhelpful; that language description should account for both what is possible and what is natural. He proposes a theory of naturalness, involving lexical priming. Finally, he explores the implications for language pedagogy. Unhelpful primings (misleading emphases in textbooks, unnatural/fabricated examples), focus on written texts, separation of grammar and lexis, error-correction, and transfer of primings from L1 to L2 all represent problematic areas for language learners. The remedy is to avoid the creation of unhelpful primings, and to create environments in which natural primings can occur.

Francis contrasts traditional descriptive grammars with data-driven grammar, and defines the characteristics of a new grammar which pays due attention to lexis and phraseology and to the meanings encoded by syntactic structures. A method (a gradual item-environment process) for compiling this grammar is outlined, and illustrated with new findings about the appositive *that*-clause. There is no constraint on the sequence in which such a grammar is compiled, and the result is a reliable specification of all major lexical items in terms of their syntactic preferences, and all grammatical structures in terms of their key lexis and phraseology. The association of semantic sets with their associated structure could lead to a grammar of typical meanings encoded by language, and recognition of untypical/foregrounded meanings.

Halliday and James start with the theoretical grammatical concept of system: a set of options with a condition of entry such that exactly one option must be chosen whenever the entry condition is satisfied (e.g. system of number; options: singular/plural; entry condition: nominal group, countable). Quantitative work in grammar depends on such a concept (because it allows set theory or other formal logic to be used as models). The study required a system identifiable by the corpus query program, of high generality, and of interest. Halliday's hypothesis of "equi" (0.5:0.5) and "skew" (0.9:0.1) systems was selected. The systems selected were polarity (negative/positive; predicted to be "skew" in favour of positive) and primary tense (restricted to non-future, therefore past/present; predicted to be "equi"). Results confirmed this: positive scored 89.85–90.75, negative 10.15–9.25; past tense scored 50.41 and present tense 49.59.

Kirk uses "micro-corpora" to explore the notion that frequencies of subordinate clauses mark Hallidayan "register" ("mode" in terms of speech and writing; and "tenor" in terms of formality and informality in both speech

and writing). Decisions are based on numerical predominance not categorical claims. Attention is drawn to corpus sizes and specific contents. Features counted can affect the results, and the degree of delicacy of the analysis. The study confirms that subordinate clauses are register markers.

Kjellmer states that natural languages are largely systematic, hence their efficiency as communicative tools. Generative grammarians have shown that it is the systematicity that allows us to encode and decode previously unencountered sentences. Large areas of the lexicon are also systematic, especially word formation. The research investigates the existence and nature of lexical gaps, system slots unfilled, for adjectives and de-adjectival nouns. The corpus contains many more adjectives than nouns, and the reasons adduced for the gaps are: non-referentiality, blocking, denominal adjectives from abstract nouns, de-adjectival adjectives, nongradable/classifying adjectives, and infrequency of the base adjective.

Part 8 Translation studies, multilingual and parallel corpora

While fully automated machine translation remains the goal of some computational linguists, corpora are finding increasing uses both as data sources within MT systems, and especially as core resources in semi-automated translation software. More attention is also being paid to multilingual corpora of various kinds used for cross-linguistic and translation research and pedagogy. **Pearson** looks at how parallel corpora might be used in translator training courses. Comparable corpora are already in use, but are insufficient. Parallel corpora specifically show how translators have overcome difficulties of translation in practice. The study looks at a small collection of popular science articles translated from English into French, focusing on a set of culture-specific references, which are known to be difficult for trainee translators and teachers, and confirms the usefulness of parallel corpora in translator training.

Frankenberg-Garcia discusses the potential usefulness of parallel concordances in second language learning. This encourages learners to explicitly compare L1 and L2 languages, which might be problematic. Navigation also presents problems, as two types of language are being compared (original texts and translated texts) as well as two languages. Should search queries be in L1 or L2? Should L2 originals and L2 translations be distinguished? The author suggests that decisions will vary in different teaching/learning contexts.

Hasselgård explores the preservation or alteration of thematic structures in translated texts, focusing on translation pairs in which word order changes cause a change in thematic perspective, or in which thematic structure is retained despite syntactic restructuring in the translation. She questions the status, especially in a translation perspective, of theme in Halliday's model:

as the first experiential element and the peg on which the message is hung. The great majority of examples showed no change of theme, suggesting that translation is a linear process.

Johansson considers different ways in which translation corpora can be used in contrastive linguistics (broadly defined). His model combines different types of corpora within the same overall framework, and each type can be used to control and supplement the other. Translation effects can be identified, and frequency distributions and stylistic preferences examined. The use of multilingual corpora enables the study of language-specific, typological and universal features. Corpora of translated texts, varying source and target languages, and learner language, varying L1 of the learners, reveal general and language-specific features. Analyses hitherto conducted on monolingual corpora need to be replicated on bilingual and multilingual corpora.

Altenberg sees the task of contrastive linguistics as establishing and describing the degree of correspondence between languages. In the past, the lack of a clear connection between *langue* and *parole*, and of relevance to language teaching, lexicography, translation, and cross-cultural communication led to disappointing results. Using bilingual and multilingual corpora for contrastive research has transformed the situation. The study of mutual correspondences between categories and items in source texts and translations reveals not only language-specific properties of the categories, but insights into the larger systems, and how the systems interact with each other and other systems. This research confirms that linguistic categories rarely show 100 per cent correspondence in translations, and suggests some reasons for this.

Aijmer asserts that it is difficult to decide when phenomena in two languages are correspondences. She uses parallel texts to see to what extent modals in English and Swedish have acquired the meaning of epistemic possibility, and how the process takes place, observing when modals are rendered by modals in translation, and when other devices are chosen. The notion of epistemic possibility is not stable across languages, and translations reveal that a variety of devices are used to express it, including modals supported or replaced by modal adverbs. The direction of translation leads to different strategies being favoured. The degree of grammaticalisation of modals is different in the two languages, and the process is gradual and affected by linguistic context.

Kenny looks at the exploitation of collocational norms in German-English translation. Target texts tend to be more conventional than source texts. She examines lexical normalisation to see whether creative compounds and collocations in German literary texts are normalised in their English translations. The corpora are small, and therefore the evidence is insubstantial. However, it is clear that translation-oriented studies of lexical creativity benefit greatly from the use of comparative corpus evidence and corpus

linguistic notions of collocation, semantic preference and prosody below and above word level.

Zanettin deals with the design and analysis of "translation-driven" corpora. The study of similar contexts and their translations, combined with statistical analysis and data manipulation, allows hypotheses to be tested on a larger scale, and tentative generalisations to be made. Design involves types of texts included, languages chosen, sampling criteria, research aims and applications, and corpus encoding. Translation-driven corpora are invaluable in descriptive and applied translation studies, allowing the study of linguistic and extra-linguistic features of translated texts on a large scale. Design criteria need to be made transparent, and alignment procedures have impacts on the findings.

Váradi suggests that grammatical morphemes are useful clues to finding translation equivalents in parallel corpora, because they form a closed set, occur frequently, have fairly fixed meanings, and have one-to-one or one-to-few relationships with elements in other languages. They can therefore make easier the task of finding word or phrase level identifications. Grammatical morphemes provide anchor points in parallel texts and can trigger local heuristic pattern-matching routines to extract translation equivalents. Many of them have stable, unambiguous equivalents on the lexical (de-contextual) level. Results can be enhanced by pre-processing (sentence alignment, shallow parsing) and using content words with similarly stable equivalences, yielding contextual equivalence data much richer than in bilingual dictionaries.

Déjean and Gaussier present a new method for the automatic extraction of bilingual lexicons from comparable corpora. They examine existing assumptions and associated algorithms, and evaluate their method using two corpora, concluding that combining their method with existing ones significantly improves the quality of the extracted lexicon. **Salkie** reflects that multilingual corpora have revived interest in contrastive linguistics, asserts that it still lacks a distinctive research programme, and makes suggestions for the programme and its underlying theoretical framework. He highlights the need for experimentation with the new corpus tools and a coherent set of working methods.

Gellerstam focuses on the global status of English, and the stream of loan words that it is donating to other languages, as well as affecting the meanings of words and phrases in them, using Swedish as an example. His ultimate concern is the role of national languages at a time when English is so dominant as the international language. The impact on industry, education, and research is clear, and the concern is about "loss of domains of usage": Swedish may cease to be used in certain domains, because it no longer has the appropriate terms. The field of activity examined is translation and the linguistic areas of impact are not just loan words, but also grammar, syntax and rhetoric. Contrastive linguistics, based on parallel corpora, is attesting facts about languages which can only be revealed by comparison.

Baroni and Bernardini recount varying attitudes to collocation, but select Bolinger's "affinities among words" as a crucial argument for studying collocation, and suggest that it may have significant impact on language theory, description, and applications. Using monolingual comparable corpora, they try to select and compare collocations across original and translated texts, and conclude that they reveal hints of systematic differences in the use of collocations.

Baker outlines the problems of studying authentic data, especially in huge (corpus) quantities, in terms of how to select the features to study, and how to interpret the findings. This requires the highly explicit elaboration of the methodology adopted. The findings can be contested by others using the same dataset, and more plausible explanations can be invoked by highlighting different parameters. She seeks to move from "low-level description to situated explanation". Frequency suggests prominent features in the data, but researchers still subjectively create the object of study, and the onus of interpretation lies with them alone. A major advantage of corpus-based work is its greater level of transparency.

Part 9 Critical discourse analysis / evaluation / stylistics / rhetoric

This section covers a wide range of diverse research interests in various language domains and genres.

Biber *et al.* comment on the neglect of spoken academic texts in previous research, which focused on research articles (not even textbooks) in science and medicine. They analyse spoken and written data at US universities, specifically a TOEFL corpus, and find strong and absolute contrasts between spoken and written registers (whatever the purpose of the text), and surprising similarities between classroom texts and conversation. Pedagogic and research conclusions highlight the wide variety of registers, from informationally dense writing to complex interactive speech. Implications are outlined for teaching, materials development, testing, university documents (advertising and administrative), and future research into lexical and rhetorical features.

Flowerdew looks at problems of description and interpretation in critical discourse analysis (CDA) under five critical claims: CDA does not deal with "facts", is reflexive, is open to multiple readings, must be plausible, and is subject to the same limitations of linguistic communication as other disciplines.

Peters records the history of the study of Australian English, and reports on corpus research into some specific features: conjunctive "like", the subjunctive, contractions, and so on, in comparison to other major varieties, confirming the Australian preference for informal stylistic options in writing as well as speech.

Geoff Thompson recommends the use of small comparable corpora in different languages (Chinese and English tourist brochures and Swiss and

English job adverts) in classroom situations to look at form, not in terms of isolated structures, but at the discourse values of lexical and structural choices in achieving communicative goals, and cultural aspects. The benefits of discourse analysis, use of corpora, and cross-linguistic comparison are discussed.

Partington examines the way in which political journalists use third party attribution of opinions hostile to the interviewee in order to change their participant status ("footing shift") and appear neutral, and the ways in which the interviewee's linguistic response can counter, attenuate or challenge such practice.

Paul Thompson looks at the use of modals in native-speaker scientific PhD theses. Academic writing textbooks over-emphasise the use of modals to express tentativeness, whereas a corpus reveals considerable variations in different disciplines and rhetorical sections.

Biber asserts that written and spoken styles diverged in the seventeenth–eighteenth centuries; that popular writing (letters, fiction, essays) moved back towards conversational style in the nineteenth–twentieth; but that recent newspaper texts also reveal evidence of innovative and demanding literate devices such as compressed noun phrases.

Teubert examines the Euro-sceptic discourse in Britain, and finds it more ritualised and distinct than in other EU countries, despite government actions speeding up integration. By contrast, German officials profess total commitment to the EU, while their actions reveal their greater reluctance. Positive opinions towards the EU seem to be limited to the intellectual elite.

Krishnamurthy uses newspaper texts, dictionaries and a large corpus to reveal discrepancies in the use of keywords connected with ethnicity. The use of "ethnic", "racial", and "tribal" is compared in relation to different parts of the world, and shows that overlapping uses of the terms, compounded by synonymous definitions in dictionaries and superficially humorous usages, serve to conceal underlying ideological attitudes, some indicative of latent racism.

Coulthard advocates the incorporation of corpora and corpus analysis techniques into forensic linguistic research. Only by establishing the "norms" in any discourse type using authentic data can we identify deviant, non-authentic, and deliberately falsified features in forensically examined texts. **Cotterill** analyses data from the O. J. Simpson trial and compares it with a large general language corpus to reveal the conflicting representations of domestic violence against women in the courtroom context, based on the different lexical realisations and semantic prosodies in the prosecution and defence arguments. **Channell** discusses evaluative, pragmatic meaning and semantic prosody in connection with selected lexical items in political, moral and aesthetic contexts, and considers the implications for theories of lexical meaning, psycholinguistic accounts of the mental lexicon, and applications such as lexicography and language pedagogy.

Part 10 Language history / historical linguistics

This section is necessarily short because many of the relevant historical texts have not yet been digitised. High costs are involved in converting to electronic form the historical texts to be studied. Many texts are rare and difficult to access, or are physically delicate and require very careful handling. Some technological processes such as scanning could damage them irreparably. Keyboarding is an expensive and labour-intensive alternative. Therefore, corpus linguistics has hitherto focused on synchronic research, where the availability of texts is less of a problem.

Nevalainen considers gender differences in the evolution of Standard English based on evidence in a corpus of Early English correspondence. The supralocalisation of regional features (before the subsequent period of overt prescriptivism and normative grammar) was influenced in some cases largely by women writers, whereas other changes can be ascribed more to their male counterparts.

Fitzmaurice studies the pragmatic meaning of modal verbs in the eighteenth century in terms of politeness strategies, and relationships between authors and their literary and political patrons. Modal choice varies (alongside other stance markers) according to register (letters, essays) and purpose of the communication, being more prevalent in humiliative texts, and reveals slow semantic-pragmatic shift in their use.

Rissanen describes the development and grammaticalisation of the preposition and conjunct "beside(s)" from Old English to Middle English, comparing evidence from corpus data and historical dictionaries. The item is shown to have developed from local concrete senses to distancing and abstract senses indicating addition, exception, or denial, and is contrasted with the development of both borrowed equivalents and other native formations. **Davies** discusses the possibilities for the investigation of syntactic and semantic change offered by different corpora of historical Spanish, because of the differences in their query language syntax.

Kytö looks at the collocational and idiomatic properties of five central verbs (*make*, *take*, *give*, *have*, *do*) in Early Modern English. Collocating object nouns are isomorphic and tend to occur in the singular. As regards idiom formation, syntactic fixity is evident especially in unmodified constructions, and specific modificational elements are examined. Variations are observed both in different periods and in various text types, but the overview confirms the progress of English from a synthetic/inflectional language to an analytic/isolating one. **Mair** utilises an utterance-based model of language change and a set of matching corpora to examine three patterns of verb complementation in current British and American usage, and interprets them against diachronic changes (evidenced in the OED quotation base) in Late Modern English grammar as a whole. Whereas divergent phonetic norms existed in the eighteenth century, grammatical variations

are a more recent phenomenon. World English is not simply converging on American norms, but subject to a complex dialectic; and grammaticalisation processes can be empirically verified even while they are in progress.

Part 11 Language teaching

This section considers the use of corpora to decide what should be taught, to investigate the output of language learners, and directly as a classroom alternative or complement to traditional methodologies and materials. **Johns** outlines data-driven language learning using discovery procedures. Noting the failure in co-ventures between pedagogy and Artificial Intelligence, he asserts that a rule-based system (which tries to encapsulate "competence") is inadequate, and a data-driven approach (accessing "performance") more appropriate. By stimulating students' questions, providing them with the relevant data, and allowing them to use their intelligence to find their own answers, this method is motivating for all levels of student, and encourages autonomy.

Meunier evaluates the pedagogical value of native and learner corpora in EFL grammar teaching. Starting from an SLA perspective of grammar, and the impact of corpus research on grammar description, she gives examples of corpus use in three areas of pedagogic application: curriculum design, reference tools, and classroom teaching. She ascribes the lack of corpus use to lack of information, reduced attention to form, and unavailability of the technology; but also warns of the limitations of corpus work, e.g. restricted context in concordances inhibits awareness of text-level features of language.

Granger first overviews learner corpus research in SLA and ELT, then discusses corpus design criteria and analytical methodologies, comparing native and learner data, and learner data of different types of students. Highlighting the advances in software, she also considers types of annotation (POS-tagging, error-tagging); research in pedagogy, curriculum and materials design, and classroom practices; and impacts on learner dictionaries, CALL programs, and web-based teaching. She advocates more integrated SLA, ELT, and NLP research, dissemination of large corpora, in-house corpora, (automated) annotation, longitudinal studies, qualitative process-oriented studies (to supplement quantitative product-oriented ones), and diversification of corpus use.

Poos and Simpson compare hedging in different academic disciplines, based on a corpus of academic spoken English, and refer to research in hedging and gender and hedging in written academic discourse. Gender is not seen to have much significance in academic speech, but differences in frequency are found between physical sciences (less hedging) and humanities (more hedging), and also in type. Such variations are important for EAP teaching, because of their interactional and social functions. However, hedging phrases

are often multifunctional, and need to be studied in the context of a multiplex speaker identity.

De Haan compares English academic writing by Dutch and English students. Dutch students have been affected by assumptions about their "near-native" English language level, and consequent reductions in teaching. Persistent and significant differences are potentially problematic and require attention: vocabulary limitations increase the use of paraphrase and adverbs, and cause syntactic differences; and national writing traditions vary.

Bernardini draws a distinction between using corpora for language descriptive insights that affect pedagogy, and using corpora directly in the teaching/learning process, and focuses on the latter, with particular reference to the teaching of LSP and translation, and suggests appropriate uses: communicative reasoning-gap activities, strategic and serendipitous learning, and for reference purposes. She emphasises discovery procedures, motivation, and autonomy, simultaneous focus on form and meaning, and the increasing availability of corpora and tools and facilities to create ad hoc corpora, but adds the caveat that corpora should form part of a range of pedagogical opportunities, which should include interaction with teachers and other learners.

Coxhead describes the development and evaluation of a new academic word list based on a corpus of academic English writing and additional to the 2000 most frequent words in English (as compiled by West in 1953). The new list contains 570 word families that account for 10 per cent of tokens in academic texts; 94 per cent of the words in the list occur in 20 out of 28 academic subject areas. The list can be used for setting vocabulary goals for EAP courses, creating teaching materials, and helping students to focus on useful vocabulary. Suitable tests need to be devised to assess whether learners know these words, and whether the words can be successfully taught and learned. Academic texts may be adapted to reduce the density of rarer unknown items, and increase exposure to more frequent items. Focused vocabulary learning yields better results than incidental learning, but should be complemented by opportunities to encounter the items in message-focused circumstances. Eighty-two per cent of the words are of Greek and Latin origin, which suggests that prefixes, suffixes and stems should also be studied. Register-specific corpora are of especial value for pedagogy, but meaning distinctions need to be researched. Subject-specific corpora and spoken corpora may also be useful. The list offers a systematic approach to academic vocabulary development.

Goutsos *et al.* look at a corpus-based approach to the research and teaching of modern Greek, highlighting linguistic relevance, lemmatisation and morphology, collocation, free variation and functional variation, word order, discourse markers, spoken data, data-driven learning, teacher training, classroom exercises, Greek in L2 and L1 teaching and textbooks, and the use of computers per se. **Mukherjee** asks which norms should be taught in ELT in

India, and how they should be implemented. He argues for a usage-based, endonormative model of Indian English, requiring corpora for its basis, to bridge the gap between Indian English use and ELT teaching in India. The Indian situation warrants, but does not yet enjoy, linguistic independence; Indian English is specifically geared to the needs of its Indian users, but there is no authoritative reference work for Indian English lexis, grammar, or style.

Part 12 Spoken language / discourse studies

One of the most significant contributions that corpora have made to linguistics has been in the field of spoken language. Although the technological problems are greater than for written texts, considerable progress has been made with generally smaller datasets. Studies of specific discourses have also developed substantially by reference to corpus collections.

Swales regards specialised micro-corpora as a viable pedagogical alternative to the general language corpora of the previous decade, especially for ESP, but urges the need for their effective and efficient use. Contrasting the high profile of corpus linguistics in Europe and the lagging behind in the USA, he warns against an over-reliance on corpora: "high frequency does not entail high pedagogical priority." Genre analysis has focused on developing a richer socio-cognitive theory, whereas corpus linguists take the concept of genre for granted. Discoursal top-down and corporist bottom-up approaches are at odds; this twin-track development may yield narrow linguistic benefits, but do not help in understanding the form–function intersection in academic speech. Poos and Simpson's findings (see Part 11 above) are skewed because there are fewer female speakers in the physical sciences than in the humanities; and terms are more fixed in science, so require less hedging. The fragmented field of corpus observations requires discoursal intuition, pedagogical priming, and a symbiosis of top-down and bottom-up approaches.

Farr investigates the linguistic devices (minimal response, non-minimal response, and simultaneous speech/interruption) used to signal engaged listenership in meetings between tutors and graduate students, quantifying them and analysing their functions. The differences are significant for the effective functioning of students in an EAP context, so the pedagogical implications are discussed: speaking-while-listening skills, variety of experience, monitoring the discipline, and corpus-based instruction, which allows critical analysis of almost any language-related issue, improving language awareness and use.

Mauranen examines English as a global lingua franca, a vehicular language used by people who do not share a native language. As non-native speakers now outnumber native speakers, native-speaker models are unsuitable and an international model is required. The first step is an accurate

description based on ELF (English as Lingua Franca) corpora, to reveal the centripetal and centrifugal forces involved. The range of variation is likely to be enormous, so preliminary research in specific contexts is necessary. An academic ELF corpus is described here, and issues of theoretical interest include discourse marking, formulaic expressions, simplification, and universally unmarked linguistic features. This will impact on the teaching of international English, technical writing, and so on.

Koester uses evidence from a corpus of workplace conversations to argue for a discourse approach to teaching communicative functions or speech acts in spoken English, focusing on giving advice and giving directives, in terms of performatives and metalanguage. The pedagogical implications for communicative competence in functional language are discussed.

Carter looks at literary language, and literariness in a range of discourses, and focuses on literary properties (partly equated with creativity) observed in everyday conversation, as opposed to well-researched discourses such as literature itself or advertising. He argues the need for social and psychological models to explain the pattern-developing and pattern-forming revealed. Research and pedagogy should continue to explore the continuities between literary and non-literary language; a fuller description of speech genres to assess degrees of interpersonality and inter-subjective accord; triggers of pleasure and linguistic marking in literary interaction and their possible culture-specific or group-specific value; and creativity in language in general, moving away from meaning in terms of truth, reference, and literalness and towards an essentially figurative view of language, involving its affective, interpersonal and bodily characteristics.

Meyer adopts a functional view of grammar, seeing language as a tool to satisfy the communicative needs of its users, rather than a formal system of rules describing the structure of isolated sentences; interest in language at discourse level is a natural concomitant. Halliday's systemic/functional grammar offers a suitable analytical framework. Comparing texts with a reference corpus provides an appropriate methodology. Speech and writing are not absolute categories, but a continuum. Registers are heterogeneous not homogenous. Academic texts seem to be more varied than other registers. The reasons may be found in cultural and ideological contexts, and different discourses conducted within the register. Variation between student writing, spoken texts, and professional writing may be due to the apprentice/imitative nature of student output. Students need to be taught not just literacy, but "powerful literacy", the ability to understand and critique the competing discourses involved in producing text.

McCarthy focuses on the relational importance of listener behaviour in "small talk", examining some of their high-frequency short-response tokens, which are superfluous to transactional needs, but of interactional value, and fulfil social functions. American and British varieties of English share many non-minimal response tokens among the most frequent 2000 words.

Their uses exhibit relational as well as feedback functions, and are often turn-initial, preceding transactional elements. They are more than just back-channel devices or discourse markers; even when freestanding, they are not turn-grabbing. They do serve to design and organise the talk, and show hearership, but additionally signal engaged listenership. Small talk appears to exist at the margin of big talk, but fulfils a significant discourse role, and forms a continuous thread in the talk, not an intermittent feature. Corpus analysis reveals its regularities, patterns, and core lexicon.

Crowdy overviews factors in spoken corpus design, on the basis of his experience with the British National Corpus, which adopted twin approaches: demographic and context-governed. The recording processes and metadata-collection are described. **Halliday** asserts that spoken language is the locus of semogenesis, the creation of meaning and the extension of meaning potential, and ponders the potential of scrutinising large corpora of spoken data as a basis for "grammatics" (the theoretical study of lexicogrammar). He sees no opposition between theory and data; observation and theory are different stages in a single enterprise of extending the boundaries of knowledge. He compares and contrasts the nature of written and spoken texts, then focuses on spoken corpora, raising problems of transcription, lack of prosodic markers, and over-transcription. The lexical focus of corpus software makes grammar harder to get at, yet speech is less word-based and more grammaticalised. More data and better measuring necessarily transform theory. In corpus linguistics, every instance carries equal weight, and the instance is a window into the system. Speech is more spontaneous and less self-conscious than writing. Speech is where systemic patterns are established and maintained, and new instantial patterns are created, which may become systemic through repetition. Speech is essentially monologic, with dialogue as an extended setting. The interaction of speech with the context of situation means that each moment both narrows down and opens up the options available at the next.