1

Introduction

1.1 Genesis of this book

This book is the result of ten years of study of the spoken language and its importance to language teaching. Initially this involved transcribing and analysing brief conversational extracts, and latterly (but with the same aim always in mind) examining large numbers of conversational extracts brought together in the CANCODE (Cambridge and Nottingham Corpus of Discourse in English) corpus in the Department of English Studies at the University of Nottingham (see 1.2 below). Over those years, I have presented and published papers and written books, sometimes of my sole authorship, often with my close colleague and co-researcher Ronald Carter (and recently also with my colleague Rebecca Hughes). Those papers and books have led me more and more into questions concerning everyday spoken language as a model for language teaching, how different types of spoken language can be classified, and what status the spoken language should have as an object of study within applied linguistics in general. That is essentially what this book is about, and the CANCODE corpus has been an invaluable tool in getting answers to (some of) these questions.

1.2 Overview

The book brings together revised versions of papers published over the period 1988–1996 and some new, previously unpublished chapters, all drawing on corpus data, occasionally quantitatively, but mostly qualitatively, for it is in the latter that I see the greatest potential for gathering useful pedagogical insights from close observation of how people ‘do’ everyday talk. In this first chapter I outline the CANCODE corpus project, upon which most of this book is based. I have tried to contextualise it with reference to other corpus projects. The chapter also takes a historical glance at the status of spoken language in language study and the
2 · 1 Introduction

teaching of languages over the centuries. There is a tendency in applied linguistics nowadays to consider anything published more than five years ago as ancient, and a historical perspective is often absent from MA course syllabuses and professional textbooks. I firmly believe that we have much to learn from our scholastic predecessors, not least in that their achieve-
ments may urge humility upon us when we are tempted to think that our ‘science’ has made massive steps forward. Chapter 1 then, concludes with a note on the contributions made by speech-act theory and discourse- and conversation-analysis to our improved understanding of spoken language, though with notes of caution expressed along the way.

Chapter 2 attempts to construct the outlines of a theory of spoken genre. In particular it focuses on the variation present in speech events which, nonetheless, still share a lot of common features. Chapter 2 also stresses the importance of looking for evidence that participants themselves are aware that they are engaging in the creation of genres, and that we are not just indulging in analytical fancies. We find such evidence partly in terms of the participants’ orientation towards past events and upcoming ones, the need for agreement among participants as to ‘where they are’ in the talk and the need to bring into effect procedures whereby events assume their final shape that we as analysts can recognise. Chapter 2 uses the CANCODE corpus matrix outlined in Chapter 1 to show how extracts display similarities at the lexicogrammatical level which correspond to higher-order features of generically-oriented activity. The chapter concludes that genre will always remain a difficult notion to pin down because social activity is prone to so much variation. What is apparent is that seen as a whole, behaviour is integrated: the transac-
tional, the interactional, the goal-orientation, the relationships among participants, and the local lexicogrammatical details all complement each other. Chapters 1 and 2 together are an attempt to sketch out a theory of spoken language that will have pedagogical relevance.

Chapter 3 raises the question of just what should and can be taught about the spoken language. I take a number of concepts from discourse analysis and conversation analysis and examine existing research to try to get answers to questions such as ‘Do these features matter?’, ‘Are they universal?’, ‘Are they likely to be transferred, and if so, under what sorts of learning conditions?’, ‘How can syllabuses and methodologies take such features into account?’. The features examined include exchanges and adjacency pairs, discourse markers, ellipsis, openings and closings, and in each case I look for the lexical and grammatical realisations of the
features and their cultural import, using CANCODE data examples to illustrate the way speakers encode the features. The chapter concludes that there is a good deal to be incorporated into teaching from the research available in the areas looked at, but that traditional, presentation-based methodologies are inadequate to the task. I propose an alternative methodology outlined by Carter and McCarthy (1995b) based on what we call ‘the three Is’, in contrast to the traditional ‘three Ps’ of presentation, practice and production.

Chapters 4 and 5 are both devoted to issues of grammar. Chapter 4 argues for a discourse-based view of grammatical choice, and uses spoken data to substantiate the claim that certain aspects of grammar are best understood when examined in context. I try to demonstrate that sometimes we need to re-think how we group items in grammatical paradigms, taking the words it, this and that as examples. I then move the discussion on to the treatment of the past perfect, and attempt to show how choice of that tense is discoursally motivated, rather than by some deterministic rules about time. The chapter continues in this vein, examining features that are differently distributed in spoken and written data, and a typically spoken feature that has long caused problems for grammarians, the English get-passive. Central to the thread of discussion in the chapter is the advocacy of the usefulness in language teaching of probabilistic statements (statements about the most likely conditions of occurrence of a form), rather than only viewing grammar as deterministic rules. I conclude that a discourse-based view of grammar, backed up by spoken data, is extremely illuminating and has direct consequences for what and how we teach in grammar classes.

Chapter 5 pre-empts two possible dangers: (a) that we may rush off and assume that everything is different in spoken grammar and that nothing we say about written language has any validity for the description and the teaching of spoken language, or (b), equally dangerously, that we should assume that descriptions of the written grammar can simply be imported wholesale into spoken grammars. Taking a discoursal framework built on the work of the late Eugene Winter, the chapter looks at how sequences of verb tense choices can ‘frame’ whole written texts and spoken episodes, and how these framing functions are found in both spoken and written. The data shows that the realisations of these functions are in some cases the same in spoken and written, but also that their realisations are sometimes crucially different. This evidence is used to support the arguments that contextual grammars as advocated by Winter are indeed
4 • 1 Introduction

very powerful descriptive instruments, and that spoken grammatical features cannot always be assumed to follow the same realisations as their written functional counterparts. It is also an exercise in showing the usefulness of direct comparisons between written and spoken texts.

Chapters 6 and 7 shift the focus to vocabulary, but still within the perspective of a spoken discourse-based approach. Chapter 6 begins by laying out some of the constraints inherent in multi-party interaction that are likely to influence vocabulary selection, and then examines various lexical features in spoken language extracts. The question of lexical density is discussed, and especially the sorts of vocabulary found in low-density, language-in-action episodes. These types of vocabulary present the teacher with challenges that traditional, content-based notions of vocabulary do not need to concern themselves with. From there, the chapter shifts to looking at how speakers take up another’s vocabulary choices, either to progress or to hold up the smooth development of the talk, and how the role of listener is just as important as that of main speaker in building the vocabulary of an interaction, especially in oral narratives. Wider issues of the differences between spoken and written vocabulary lists generated from corpora are then examined, and the sorts of problems frequency-counts raise for teaching are touched upon. The chapter concludes that training learners’ observational powers might be just as useful as pumping them with large numbers of new words.

Chapter 7 is about idioms, a subject most language teachers can identify with as part of the stock-in-trade of vocabulary teaching, especially at higher levels of proficiency, and yet one that has rarely been dealt with from a discourse perspective, and one where available information on usage in everyday spoken language is scant. After discussing problems of definition, I offer data from spoken narratives to support the claim that idiom selection is not random, but plays a key evaluative role in storytelling. The chapter then casts its net wider, with examples from discourses where speakers are commenting on aspects of their world in general; once again, idioms have an important evaluative role to play. The data shows that speakers often ‘unpack’ the literal meanings buried in idioms and create extended puns and images that thread through the text. Idioms also have a role in reinforcing cultural membership. The chapter ends with a discussion of the problems of teaching idioms, and the kinds of exercises and contexts that can be exploited while remaining faithful to a discourse-oriented approach.

The last chapter of the book returns to a grammatical area that usually
has its place in any language-learning syllabus: reported speech. We begin by comparing two extremes in speech-reporting: classic literature and everyday conversational storytelling. The conversational reporting looks at first glance to be rather impoverished when compared with great literature, but I attempt to show its richness and flexibility. I then look at other spoken reporting strategies, such as past continuous reporting verbs and the narrative historical present. With the quantified evidence from the CANCODE corpus, the chapter then compares what is frequent in the corpus with what is usually found in literary reporting, and shows that the corpus examples are very varied, serving a variety of strategic functions. The chapter ends with a general set of recommendations for teaching speech reporting, as well as a general conclusion on the book as a whole. (There is a glossary of terms on pages 176–81.)

1.3 The CANCODE corpus and its context

The CANCODE (Cambridge and Nottingham Corpus of Discourse in English) project was established in the Department of English Studies at the University of Nottingham, UK, with the help of a generous research grant from Cambridge University Press, the publisher of the present work and (to date) of two other works (Carter and McCarthy 1997a, and McCarthy and Carter 1997a) which draw on CANCODE material. A complete list of publications based on CANCODE is given at the end of this book. Cambridge University Press’s support enabled Ronald Carter, me and our research assistant Jean Hudson to assemble an initial corpus of 1,061,274 running words of informal spoken English, completed in 1996. Further support from Cambridge University Press is, at the time of writing, enabling us (with the help of our new additional assistant, Julia Harrison) to expand the corpus to five million words.

It must seem, to a profession dazzled by ever larger language corpora, a rather puny enterprise to be working with only a million (or even five million) words. Corpora now regularly consist of hundreds of millions of words and the race to be first to hit a billion is undoubtedly one that will run in tandem with the tick of the clock towards the new millennium. Technology seems boundless and gigantic corpora can now be marshalled with an ease that would have been considered pure science fiction when I entered the ELT profession in 1966. CANCODE, therefore, is numerically small by today’s standards. It is also relatively modest even by the standards established some time ago by pioneers in corpus linguistics.
6 · 1 Introduction

It would be unfair, however, to judge CANCODE’s size only in relation to huge written or mixed written/spoken corpora. More important is to consider it in relation to the development of spoken corpora in general, and to the problems that make spoken corpora much more difficult to assemble than written ones. Spoken corpora are not new: anthropologists and dialectologists have long used tape-recorded data as a major source of evidence (Biber 1990), and such work continues (for example, the Northern Ireland spoken corpus described by Kirk 1992). Some of the earliest spoken investigations were carried out within the study of child language acquisition (an example of this is the child-language word-frequency analyses described in Beier, Starkweather and Miller 1967). A notable early spoken corpus project of the kind that has since become quite common was the Oral Vocabulary of the Australian Worker (OVAW) (Schonell et al. 1956 gives a full account of the data and its collection). The OVAW corpus consisted of some 500,000 words of spoken language and is still very useful for anyone interested in idiomatic words and phrases in speech (e.g. Schonell et al. 1956: p. 67), a subject I pursue further in Chapter 7 of this book using CANCODE data. The OVAW also records the ubiquitous discourse markers found in everyday spoken language. A decade after OVAW, the Davis-Howes Count of Spoken English (Howes 1966) in the USA brought together half a million words of interviews with university students and hospital patients, and produced some interesting statistics for spoken usage.

Also in the 1960s a spoken-word count for Russian was published (Vakar 1966), which was small (based only on 10,000 words taken from drama texts), but which offered useful statistics about text coverage of high-frequency words. The University of Leuven Drama Corpus (approximately one million words from contemporary plays; see Engels 1988 for details) continued this tradition of using drama texts as a model for the spoken language (see also the reference to Vanrespaille 1991 in Chapter 4). Further evidence that literary (or at least written-fiction-based) corpora can also be useful for the comparative study of spoken language may be seen in the work of researchers investigating the 1.5 million-word TOSCA corpus at the University of Nijmegen (e.g. see Oostdijk’s 1990 study of fictional dialogue, based on TOSCA).

Meanwhile, in Great Britain, the half-million-word spoken segment of the London-Lund corpus (itself a half of the one-million-word spoken/written Survey of English Usage (SEU) at University College London; see Svartvik 1990), and the conversational transcripts available in Svartvik
and Quirk (1980) have been instrumental in some very important investigations of the vocabulary of everyday spoken English (see McCarthy and Carter 1997a for further examples). Nowadays, as noted above, gigantic corpus projects such as the COBUILD Bank of English (see Moon 1997 for a recent description of the project) and the British National Corpus (see Crowdy 1993 and Rundell 1995a and 1995b for details of design and content) contain considerable amounts of spoken English data, including broadcast speech as well as everyday unrehearsed conversation. Broadcast data of many different types form the basis of the British English Lancaster/IBM spoken corpus (see Knowles 1990). In the USA, work by Chafe and his colleagues, initially based on the British London-Lund spoken corpus design (Chafe, Du Bois and Thompson 1991), has developed into larger enterprises such as the five-million-word Longman Spoken American Corpus (see Stern 1997), which Biber and others are investigating to great effect, and which is planned to feed directly into language-learning resources.

Australian English has been subjected to corpus analysis in the Macquarie University corpus project (see Collins and Peters 1988), and by Eggins and Slade (1997), who look at everyday conversational activities such as gossiping. Also important is the ICE (International Corpus of English) project, which plans to bring together parallel corpora of one million words from 18 different countries where English is either the main language or an official language. The samples in the ICE corpus include 300 spoken texts, although these include many scripted samples, and broadcast interviews and discussions, with only 90 samples being face-to-face informal conversations (see Nelson 1996; also Fang 1995). Spoken corpora have, therefore, come of age, and many lessons have been learnt along the way. What all the projects mentioned so far have contributed to our understanding of spoken language and of corpora in general is massive. Amongst other things, we have gained a better understanding of the types of talk people engage in, instead of simply assuming that the text topologies of written language applied the same to spoken. We have gained a great deal of experience about transcription of speech (see 1.4 and 1.5 below). We have also learnt the advantages (e.g. time/cost) and disadvantages (e.g. lack of naturalness) of spoken corpora taken from broadcast or drama-text sources. Technology has enabled huge improvements in sound quality and unobtrusiveness of equipment. Last but not least, being aware of previous corpus projects helps to prevent the continual re-invention of the wheel, and builds up a body of
8 • 1 Introduction

evidence against which new corpora can be evaluated. Things have come a long way since the early recordings of dialectologists and anthropologists, but the design of spoken corpora is still often opportunistic (‘get whatever data you can’). The CANCODE corpus has tried to avoid opportunism, and to follow design principles that will make its material maximally useful to teachers, pedagogically-oriented researchers and materials writers (see 1.4 below).

1.4 The CANCODE corpus and generic features

In setting up CANCODE, the research team decided to attempt to cover as many useful speech-types (useful in terms of language learners’ perceived needs) as was feasible. However, the immediate problem was that no satisfactory classification of ‘text types’ for spoken language was available to parallel existing text typologies for written language. Chapter 2 discusses this problem within the framework of genre theory and demonstrates the theoretical stance the CANCODE team finally adopted. Here I shall outline the more practical steps that were necessary for the operationalisation of the notion of speech-types in the corpus design.

In the gathering of spoken data for corpora two approaches seem prevalent.2 The first of these may be termed the ‘demographic’ approach, where a population of speakers is targeted and where that population records its spoken output over a given period of time. Biber (1993) stresses that the targeting of the population is much more important than sampling size. A well-chosen population sample would certainly seem, intuitively, to generate a qualitatively better corpus than mere opportunistic sampling or dumping of huge amounts of undifferentiated text simply to compete in the corpus ‘numbers game’ that is gripping the profession at the time of writing. Crowdy (1993) also argues for a demographic approach, which was used by the British National Corpus researchers. The other approach may be termed the ‘genre’ approach, in that it attempts to target not only a population of speakers but particular environments and contexts in which spoken language is produced. This approach does not simply rely on a pre-ordained notion of what a speech ‘text’ is. As Atkins et al (1992) have argued, there are various ways of defining textual boundaries in speech (e.g. the moment interlocutors come together or part, or the marking of opening and closing features linguistically). The genre approach tries to seek a balance between speaker, environment, context and recurrent features. It has the advan-
1.4 The CANCODE corpus and generic features

tage that the corpus can be analysed from different perspectives (e.g. types of speakers, emergent text types, situation types, etc.). It has the disadvantage that genre is an ill-defined notion in the study of spoken language in general (see Chapter 2). In the genre approach, decisions have to be made about situational/contextual types as well as population types, and these decisions are by no means straightforward. The CANCODE project, which is based on a genre approach, confronted the problem of generic coverage by attempting to control contextual variables of different kinds in the collection of data. The data once collected could then be examined for ‘episodes’ (or linguistically marked speech events) that displayed similar linguistic patterning at both the global levels and the local levels. This will be exemplified below. The resultant model provides a proposed classification that brings out the commonality of spoken episodes; it does so in a way that offers the possibility of linking their contextual and social features directly with the lexicogrammatical ‘nuts-and-bolts’ of their step-by-step creation. The model eschews categories such as ‘written-to-be-spoken’ or ‘rehearsed spoken’, which have traditionally informed the study of variation in speech and writing (see Crystal 1995 for a good recent discussion). This is because (a) it is very difficult to know whether something is one thing or another (e.g. a radio interview, or a university tutorial) and (b) the CANCODE team decided to focus on, wherever possible, unrehearsed, non-formal talk.3

Five broad contexts for data collection based on the type of relationship among participants were identified (principally by the team’s corpus manager, Jean Hudson):

- Transactional
- Professional
- Pedagogical
- Socialising
- Intimate

Transaction relationships were defined as those where speakers display needs or imperatives and move towards satisfying those needs in a goal-oriented fashion outside of the contexts of professional, socialising or intimate relationships. A clear example would be day-to-day service encounters in shops, restaurants, etc., between servers and customers, transacting goods, information or services. Professional relationships are displayed in talk between professional colleagues in professional situations (e.g. informal company meetings, staff meetings, desk-to-desk talk).
10 · 1 Introduction

Pedagogical relations are those between teachers and their students and student–student (e.g. informal tutorial conversations, pair- and group-work). Socialising relations accord with social or cultural activities entered upon by participants but not in professional or intimate settings (e.g. a group of friends preparing a party, talking with a stranger on a train). Socialising is thus one of the most common categories, covering much of our day-to-day activity. Intimate relations pertain between family members or close friends in private, non-professional settings.

For each of these categories, three typical goal-types were posited:

- Provision of information
- Collaborative tasks
- Collaborative ideas

Provision of information is predominantly uni-directional, with one party imparting information to others. The role of information-giver may, of course, rotate among participants, but the dominant motivation for the talk is information-giving (e.g. an enquiry at a tourist information office). Collaborative tasks show speakers interacting with their physical environment while talking (e.g. two people packing a car prior to a journey). Collaborative ideas are concerned with the interactive sharing of thoughts, judgements, opinions and attitudes. Just as the context-types are broad and refer to predominant rather than exclusive traits, so too are the goal-types.

Some examples of how the categories were operationalised are given in figure 1:

<table>
<thead>
<tr>
<th>Context-type</th>
<th>Goal-type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transactional</td>
<td>Information provision</td>
<td>Tourist information office requests for information</td>
</tr>
<tr>
<td>Professional</td>
<td>Information provision</td>
<td>Company sales conference, informal informational talks</td>
</tr>
<tr>
<td>Pedagogical</td>
<td>Collaborative ideas</td>
<td>University small-group tutorial</td>
</tr>
<tr>
<td>Socialising</td>
<td>Collaborative task</td>
<td>Relatives and friends preparing food for a party</td>
</tr>
<tr>
<td>Intimate</td>
<td>Collaborative ideas</td>
<td>Mother and daughter discuss family matters</td>
</tr>
</tbody>
</table>

Figure 1: Examples of operationalised categories